St. Petersburg University
Graduate School of Management

Master in Management Program

MARKET BASKET VISUALIZATION FOR HYPERMARKETS WITH THE USE OF BIG DATA ANALYTICS

Master's Thesis by the 2nd year student
Concentration — Management General Track
Kseniia Krokha

Research advisor: Professor Tatiana A. Gavrilova

St. Petersburg

2018

LIST OF CONTENT

ЗАЯВЛЕНИЕ О САМОСТОЯТЕЛЬНОМ ХАРАКТЕРЕ ВЫПОЛНЕНИЯ ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ

Я, Кроха Ксения Сергеевна, студент второго курса магистратуры направления «Менеджмент», заявляю, что в моей магистерской диссертации на тему « Визуализация продуктовых корзин для гипермаркетов с использованием Big Data аналитики», представленной в службу обеспечения программ магистратуры для последующей передачи в государственную аттестационную комиссию для публичной защиты, не содержится элементов плагиата.

Все прямые заимствования из печатных и электронных источников, а также из защищенных ранее выпускных квалификационных работ, кандидатских и докторских диссертаций имеют соответствующие ссылки.

Мне известно содержание п. 9.7.1 Правил обучения по основным образовательным программам высшего и среднего профессионального образования в СПбГУ о том, что «ВКР выполняется индивидуально каждым студентом под руководством назначенного ему научного руководителя», и п. 51 Устава федерального государственного бюджетного образовательного учреждения высшего образования «Санкт-Петербургский государственный университет» о том, что «студент подлежит отчислению из Санкт-Петербургского университета за представление курсовой или выпускной квалификационной работы, выполненной другим лицом (лицами)».

(Подпись студента)

25.05.2018 (Дата)

STATEMENT ABOUT THE INDEPENDENT CHARACTER OF
THE MASTER THESIS

I, Kseniia Krokha, (second) year master student, program «Management», state that my master thesis on the topic «Market Basket Visualization for Hypermarkets with the Use of Big Data Analytics», which is presented to the Master Office to be submitted to the Official Defense Committee for the public defense, does not contain any elements of plagiarism.

All direct borrowings from printed and electronic sources, as well as from master theses, PhD and doctorate theses which were defended earlier, have appropriate references.

I am aware that according to paragraph 9.7.1. of Guidelines for instruction in major curriculum programs of higher and secondary professional education at St.Petersburg University «A master thesis must be completed by each of the degree candidates individually under the supervision of his or her advisor», and according to paragraph 51 of Charter of the Federal State Institution of Higher Education Saint-Petersburg State University «a student can be expelled from St.Petersburg University for submitting of the course or graduation qualification work developed by other person (persons)».

(Student's signature)

25.05.2018 (Date)

АННОТАЦИЯ

Автор	Ксения Кроха
Название ВКР	Визуализация продуктовых корзин для гипермаркетов с использованием Big Data аналитики
Образовательная программа	Менеджмент
Направление подготовки	Менеджмент
Год	2018
Научный руководитель	Гаврилова Татьяна Альбертовна
Описание цели, задач и основных результатов	Проблема аналитики больших массивов данных актуальна в разных областях бизнеса, и индустрия ритейла не исключение. В гипермаркетах количество транзакций нарастает с каждой минутой, что осложняет процесс анализа данных. Целью данного исследования является создание новой модели анализа с помощью экспериментальных SQL запросов (без использования кластерного анализа), работающих на Big Data, и создание наглядных визуальных моделей потребительской корзины. Программные продукты, используемые для решения поставленной задачи: Oracle Data Miner и Oracle Data Visualization. Глубинные интервью также являются важной частью исследования на всех этапах. В результате работы были получены не только пары товаров, но целые продуктовые корзины. Всего было получено семь разных видов корзин потребителей, которые невозможно было бы идентифицировать без визуальных моделей. Созданные визуализации успешно внедрены в работу одной из крупных сетей гипермаркетов.
Ключевые слова	Аналитика больших данных, визуализация больших данных, визуальные модели, большие данные в ритейле, анализ продуктовых корзин, визуализация потребительской корзины

ABSTRACT

Master Student's Name	Kseniia Krokha
Master Thesis Title	Market Basket Visualization for Hypermarkets with the use of Big Data Analytics
Educational Program	Management (MiM)
Main field of study	Management
Year	2018
Academic Advisor’s Name	Professor Tatiana A. Gavrilova
Description of the goal, tasks and main results	The issue of Big Data analysis is crucial in different fields, and retail is not an exception. Increasing amount of transactions every minute in hypermarkets make data hard to analyze and extract value from it for the business. There is a simple tool for identification of customer behavior, called market basket analysis, which is based on association rules. This method helps to identify cross-selling pairs of products. The problem is that in the existing researches market basket analysis is conducted with the use of cluster analysis, which is not applicable for Big Data. Apart from that, the scientific works explored are lacking clear visualization models that could be applied by managers for decision-making process. The purpose of this research is to solve these issues and to create a new model with the use of experimental SQL Queries, applicable for Big Data and visualize the results effectively. All the steps are conducted in Oracle Data Miner and Oracle Data Visualization and supported by in-depth interviews with managers in retail. Not only cross-selling pairs of products, but whole baskets were generated. Overall, seven different market baskets structures were obtained and visualized. These models are successfully implemented in one of the biggest retail chains in Russia.
Keywords	Big Data, Big Data analysis, Big Data visualization, visualization models, market basket analysis, Big Data in retail, market basket visualization, association rules

Introduction

Nowadays it is impossible to imagine life without technology, endless information flows. Regular data processing applications in many cases are not able to deal with extremely large data sets, and these huge data sets are called Big Data. Enormous data sets keep increasing, developing rapidly and it is becoming more and more difficult to structure data, analyze it and get value from it.

Exploration of Big Data is also a key issue for business and retail industry is not an exception. Huge amount of transactions take place every hour in hypermarkets. At first sight, information stated in regular receipts might not seam meaningful. It contains time when the purchases were made, number of cash registers in the store, the list of products with prices, amount of goods bought, the name of cashier, etc. Petabytes of this data is stored in databases of big retailers. However, exploration of this data will give an opportunity to find hidden trends, analyze customer behavior, improve merchandising and create many other improvements. Based on analyzed data, enhancements in different fields will result in an increase of profit.

Market Basket Analysis is one of the approaches for understanding of customer behavior. However, the goal of this method is limited; it is devoted to search of combinations of products that frequently correlate in transactions. Probably, this method is underestimated and could be used for wider range of purposes. Is it possible to expand the borders of this approach and use it as a tool for segmentation of clients and profit growth? This is a topic to be explored.

Big data analysts can extract precious information for business from “raw” data, however, it is more important to “communicate” this data to managers, marketers and other specialist who will take decisions about changes and further improvements. That is why visualization plays one of the key roles in data interpretation. As visualization techniques are effective for communication between specialists with different background, it should also simplify interpretation of Big Data. This is also an issue to be researched deeper.

The purpose of this master thesis is to visualize market baskets based on the analysis of Big Data and to identify the role of visualization techniques in data analysis. Data Mining methodology is the base of this research.

The research questions of this research can be formulated as follows:

- What are the most common market basket structures?
- How does visualization improve data interpretation?

First chapter is focused on definitions and description of Big data, Visualization, market basket analysis and main trends in these fields.

Second chapter is devoted to methodology of this research. Data mining techniques, in-depth interviews and expert opinion are the main methodologies that are used and described in more details.

Third Chapter represents the stage of data transformation with SQL queries. After that results of visualization models are shown and interpreted for further segmentation of customers of the researched company. Managerial implications and conclusions are drawn in the final chapter as well.

Chapter 1. Big Data Visualization in retail – theoretical overview

1.1What is Big Data? Theory, implication and main trends

When people started working with enormous amount of data?

Let us skip ancient times, libraries of parchment planks and move to 20th century when technology started developing rapidly. As soon as, signals turned digital, conversions of traditional libraries into machine-readable files happened, first Optical Character Recognition tools were launched. [Coyle, 2006]. Digital format of signals gave an opportunity to make data structured; and at the beginning of the "digital era" the amount of data was not out of control and it could be processed, however, at that period of time the importance of data collection and analysis was not evaluated and the appropriate tools were not invented. Nowadays scientists work on different tools for manipulations with Big Data but in modern world the amount of data is enormous, it is expanding and changing constantly and analysis of Big Data is getting more complex and the outcome is not always reliable. This is a field to be expanded in the future.

What is Big Data?

It is impossible to conduct the research without understanding of main concepts. First of all, it is important to understand what data is, what is the difference between data and information. Even though "data" and "information" are tied together, there is a significant difference in these terms and understanding of it is crucial for comprehension of further terminology in this research. Data is "raw", it has to be processed; information is the final form of processed, organized and structured data, which is presented in a way useful for receiver.

Big data is a term for massive data sets having large, more varied and complex structure with the difficulties of storing, analyzing and visualizing for further processes or results. [Seref Sagiroglu, Duygu Sinanc, 2013].

In other words, this is a term for large volume of data, both structured and unstructured. This is the reason why it is so difficult to manipulate it and people are just on the way of understanding what it really is and how beneficial right usage of Big Data can be. The term "Big Data" is relatively new (the definition was given in 1990s by John R. Mashey, famous computer scientist in USA), however, people gathered, interpreted, saved huge amount of information ages ago. Later on, in 2001, Business analyst of the company "META Group" (nowadays the name of the company is "Gartner") Douglas Laney gave a more comprehensive overview of "Big Data" term, so-called "3Vs" model. It states that amount of data is increasing regularly (expanding Volume), the speed of data is also growing (Velocity) and the variety of data types and sources is huge and it keeps booming (Variety). [Laney, Douglas, 2001].

Interestingly, that the “3Vs” model is still widely spread and used in different fields and industries. [De Mauro, Andrea; Greco, Marco; Grimaldi, Michele, 2016].

In 2012 the definition of “Big Data” was updated again by “META Group”, this is exact citation: “Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization”. [Beyer, Mark, 2011].

Some experts believe that “3Vs” model is not applicable anymore and Big Data can be described by “5Vs” model. These extra “Vs” are Variability (incompatibility of data leads to complication of management and any other manipulation processes with Big Data) and Veracity (constant changes and variety of data influence the quality of captured data, as a result, the analyses process of data becomes more difficult and the results are not sufficient). Scientists keep arguing about relevance of extra “Vs” because they are linked to “analyses stage” of Big Data, it should not be necessarily included in the definition because analysis is manipulation with data. However, complexity of analysis of Big Data itself emphasizes peculiarity of Big Data (the data flow is so enormous, so quick, so different that analysis of this data is so hard, due to inconsistency and quality of captured data). Opinions of experts vary about the type of the model “3Vs” or “5Vs” but the basic “3Vs” remain the same: Volume, Variety, Velocity. [Hilbert, Martin, 2015].

Simple example: how Big Data works

In 2009 the Centers for Disease Control and Prevention in the United States got warning messages about new type of flu. Patients felt weakness for several days but did not visit doctors before hard and dangerous symptoms appeared. When the grave stage of disease started, it was difficult to heal it and patients got serious complications. In many cases this lead to deaths. This epidemic spread quickly and the two-week lag (when patients delayed consultations with doctors) was crucial in this situation.

In a couple of weeks Google published a paper where authors explained “how Google could “predict” flu in the United States at that period of time, not just nationally, but down to specific regions and even states”. The key to “prediction” was not based just on regular searches of users, for example “medicine for cough and cold”, Google has the way more data to work with. Google analyzed the regions where searches were requested, compared data about flu and states with those that The centers for disease Control and Prevention had for previous years and identified a huge number of mathematical calculations and models. These models showed a strong correlation between “prediction of Google” and the official figures nationwide.

This “prediction” of Google was an effective processing of Big Data that gave valuable information for the healthcare sector and the government. This is a great example of “Big data in action”. [Viktor Mayer-Schönberger, Kenneth Cukier, 2013].

How “big” is Big Data?

Information society keeps “consuming” technological goods – organizations buy developed technology systems for their offices, people buy new gadgets and take their cellphones everywhere. To give an evident example, people upload more than 10 million photos every hour on Facebook, over 500 terabytes of data is processed by this social media every 30 minutes. Hard to imagine that, especially knowing that this website was launched not so long ago, in 2004.

IBM provides us with impressive statistics: daily 2,5 quintillion bytes of data is created by users. This is a good example for understanding of how rapidly data is expanding – interpreting previous statement – 90% of the existing data in the world was created for the last two years.

Why analysis of Big Data is important?

Rapid development of technology leads to expansion of Big Data, active digitization. It’s getting more and more complicated to control, process, share, create and utilize such amounts of data. As a result, people have lots of information of all types and structures, however, uncontrollable and unanalyzed data includes many contradictions, inaccuracy and this data loses value.

Information overflow is not that obvious but it exists and it is becoming a problem. “Not only is the world awash with more information than ever before, but that information is growing faster”. As time goes by, people may know only “what” but not “why”, correlation between simple concepts may be lost and not understandable. This will have a negative influence on decision-making processes, not only on a global scale but also in very basic things. People have to develop new tools that should outrun the speed of data expansion. “The real revolution is not in the machines that that calculate data but in data itself and how we use it. [Viktor Mayer-Schönberger,Kenneth Cukier, 2013].

Even though people are at the beginning of the way to understanding of Big Data and methods of manipulation with it, information society relies on it daily. For example, “autocorrect” is based on what the person types, it allows to add new words to the dictionary, and based on the updated version of the particular user, it offers new words (together with those that this particular user conducted himself). And this is just the beginning. The biggest problem is to convert Big Data into usable and valuable information. If people boost development in understanding, control and

prediction of data, new era of technology will start. Many aspects of the world will be replaced by computer systems.

Where Big Data can be used?

Thinking about Big Data, the first thing you imagine – technology, something about IT. Google, Amazon, eBay, social media – every day petabytes of data is processed. This is right that information technology is one of major fields where Big Data is actively used but computer science is not the only sphere. Big Data is playing a significant role in different aspects of life.

Sports

Big Data can be helpful for analysis of methods of training, competitor's advantages. Information about competitions gives an opportunity to bet and to predict winners, follow results, track development of athletes and this or that kind of sport in general.

Science

For many experiments of different types sensors are widely spread. A famous example is Large Hadron Collider, sensors of which replicated 200 petabytes of data during the experiment in 2012.

Big Data in organizations

According to report of McKinsey global institute (2011) "Big Data: the next frontier for innovation, competition, and productivity", those companies that will successfully analyze big sets of data and get value from this information will become leaders. McKinsey Global institute conducted a research in healthcare, manufacturing, retail in the USA, public sector in Europe and personal-located data globally. Conclusions were impressive: "Big data can generate value in each". Government, users of services, companies can save billions of dollars just by processing Big Data efficiently, extracting value from analysis. Companies have to take Big Data seriously because this will be the key to competitive advantage and innovation. Talent management will also generate employees that will handle Big Data analysis. This trend already exists worldwide but in some time only those companies will be successful that "capture value from deep and up-to-real-time information".

Healthcare

Standardized medical terms, repositories of information about patients, automated reporting of data about every visit to the doctor, diseases, predictive analytics and many other things are now possible due to Big Data. However, there is no good without evil. Human health is extremely valuable and mistakes are not allowed. Too much data leads to unreliability of information. Missing data should be fulfilled but without damage of existing quality information. Scientist need to modify intelligent tools regularly, in order to take control over the information. Healthcare sector is getting more and more electronic and the problem of Big Data is threatening – it is difficult to use it but the situation is not hopeless.

Threats are obvious but the reason for this is lack of outstanding tools for solving these problems, but opportunities are bright if people keep developing and researching Big Data processing tools. Healthcare is not able to heal all varieties of diseases possible (for now), even though the medicine is progressive in modern world. Massifs of information extracted from patient registries, description of their illnesses and symptoms might be helpful for filling in the gaps of unexplored.

Retail

Amount of data in retail grows every single day; hypermarkets generate petabytes of data every hour. A lot of important information is hidden in this data, which could give perfect overview about customer behavior, location, transactions; however the main problem lies in analysis and transformation of this data. [Bradlow, Gangwar, 2017]

Big retailers hire actively data analysts, in order to explore existing and endlessly increasing data, however, there are many issues for effective analysis on the way, such as unstructured data, limitations of information systems and other.

Anyway, exploration and exploitation of Big Data in retail is worth trying because Big Data analysis gives opportunities to segment customers, understand their purchasing habits, adopt to the changing demand, find perfect location for new stores, implement marketing tools more effectively, develop new products, track clients' journey across touch-points, control prices and extract other valuable information from structured data. [Gutierrez, 2017]

Dimensions of Big Data in Retail are represented in the figure below (Figure 1).

Fig. 1 Dimensions of Big Data in Retail

Retailers have a huge field for exploration of Big Data in different dimensions. Improvements in each of the areas will lead to perfect understanding of the customers, locations, marketing, which will result into strong advantage over competitors that have not realized the importance of Big Data analytics.

1.2 Big Data Visualization tools. Importance of data communication

What is Big Data Visualization?

Data Visualization is the way more "young" term than Big Data. Data Visualization is the visual representation of Data, in other words, it is communication of units of information in a clear form through graphical tools (such as plots, graphs, etc.). The term of information visualization is widely spread as well (and is considered as a synonym of data visualization); however in this case some clarifications are needed. In this situation, definitions of "data" and "information" are helpful for understanding of the difference in this case. Information visualization is about "large-scale collections of non-numerical information". Information visualization is connected with information united in an approximate scheme (it has structure, even if it so not so obvious). Data visualization, however, can not be easily transformed into schematic form but it can be abstracted schematically. [Michael Friendly, 2008].

Visualization in Business

Importance of computer-rendered visualization tools for decision-making in business used to be underestimated, that is why this phenomenon is relatively new. However, now, in the age of Big Data, successful businesses use actively visualization and graph analysis, in order to transform complex and interconnected data into meaningful visual models that can show hidden relations valuable for managerial decisions and represent business opportunities. Visualization tools are applicable in different areas of business, such as process optimization, financial analysis, risk and influence analysis and market basket analysis. [Brath, Jonker, 2015]

Big Data processing tools for visualization

Human brains comprehend visualized information easier than data presented in any other form. The task to analyze enormous amount of data in a static form is a serious challenge. Traditional data visualization tools are not applicable for enormous amount of information. If all the points of Big Data are going to be visualized, it will be impossible to comprehend this type of displayed information. If data is going to be reduced significantly, some important elements of data can be lost or unusual outliers might not be taken in a consideration. Apart from that, the image perception is also limited by physical perception of humans. Another challenge is "visual noise" – some of the objects might be so related with each other that it's impossible to separate them or remove them both completely. This is the main problem of Big Data Visualization. [Qunchao Fu, Cong Wang, 2014].

The visualization tool should be able to provide the user with interactive visualization with as low latency as possible. [Syed Mohd Ali, Rakesh Kumar Lenka, 2016].

As it has already been mentioned, traditional tools for visualization are not appropriate for Big Data in modern conditions. Nowadays, such tools are used as Tableau, Microsoft Power BI, Gephy, Plotly, Oracle Data Visualizer. Visualization with the use of these tools is effective, and the functions of such IT products allow to work with huge data flows.

1.3 Market Basket Analysis

What is market basket analysis?

Market basket analysis is one of the tools for understanding of purchasing behavior of consumers, which is widely used by retailers. In other words, market basket analysis is a special technique that helps to find combinations of products that frequently correlate in transactions. This means that if customers buy certain group of items, most likely they will add to cart another group of items; the point is to find frequent combinations between groups of items [Albionresearch.com. 2018].

Is it implementable outside retail?

At first sight Market Basket analysis might look like a specific approach of analysis for retail industry. However, it can be used in different fields because this method estimates multiple things done by customers in close proximity.

- Hotel Business: if the customer books a hotel by credit card or pays for the room in the hotel, it is possible to track client's traveling preference and predict what service this consumer will purchase next (SPA-procedures, car rent, tickets to the museum, etc.)
- Telecommunication: it is possible to develop special "packages" for customers based on their preferences (offer extra TV-channels, unlimited packages for calls to some cities, packages with different Internet traffic, ISDN, etc.). The goal is to offer more by choosing what is interesting to this particular client
- Banking: if the customer has a deposit in the bank, it is possible to offer insurance services in case of robbery near cash machine and other additional services. Moreover, in cooperation with retail chains, it is possible to track transactions and predict what kind of service the client will likely need next (car loans, investment services, etc.)
- Market basket analysis is a good solution for fraud identification. Unusual manipulations with card transactions, strange combinations of insurance claims could be a signal for deception.
- Medicine: Analysis of combinations of treatment fixed in digital medical records could help to identify complications, possible allergic reactions, etc.

These are just some examples of possible implementation of market basket analysis approach.

Advantages and limitations of this approach

Market Basket analysis is attractive to analysts because results are based on association rules, which are clear and simple. Association rule is a technique that is meant to find correlations, associations and using the criteria support and confidence to identify the most important relationships in data sets. Support shows the frequency of the items that appear in explored data sets. Confidence represents the amount of times the if-then statements in the given data sets have been selected to be true. [SearchBusinessAnalytics, 2018]

Association rules are easy for comprehension and understanding, however, it does not mean that the outputs are always meaningful. In order to obtain valuable results, generated rules should be interpreted carefully [Berry, Linoff, 1997]. Examples of possible interpretations are shown in the figure below (Figure 2).

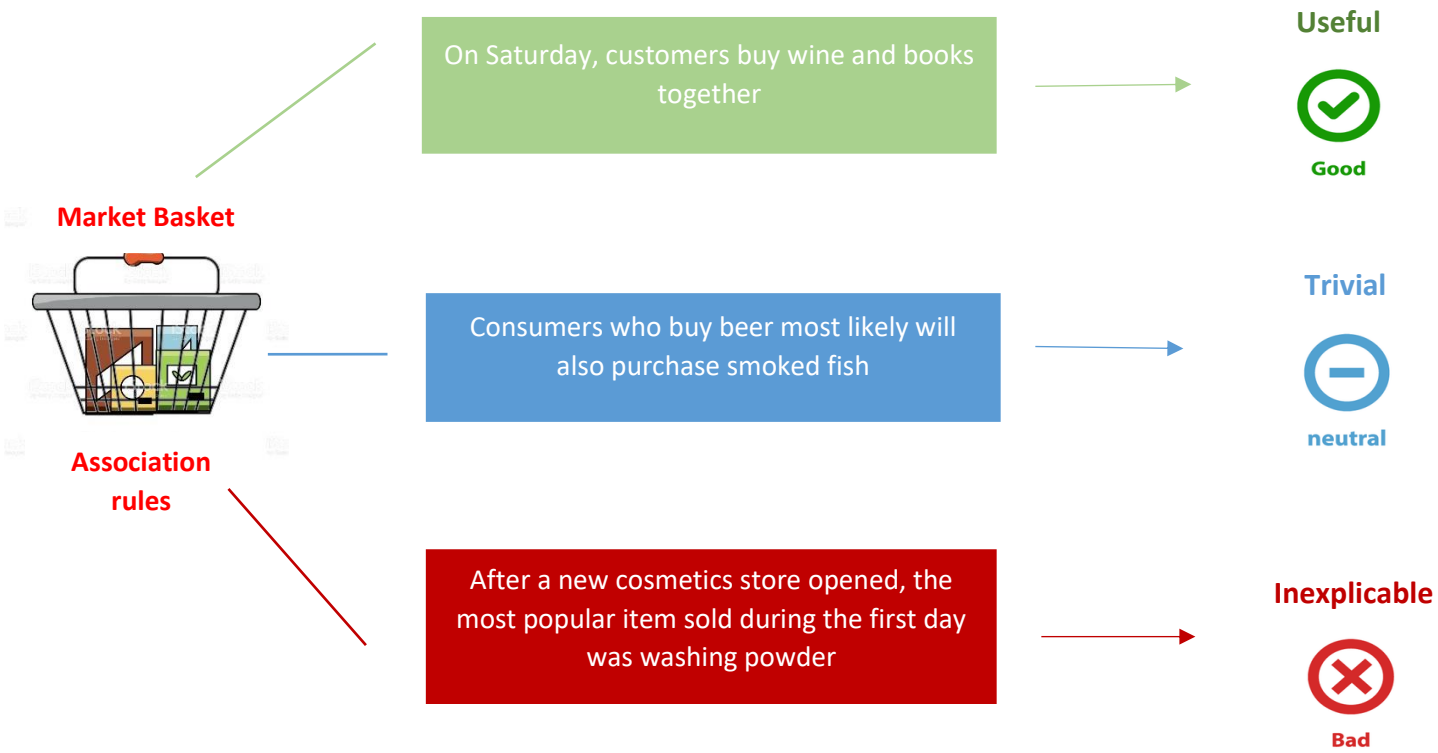

Fig. 2 Examples of interpretation of association rules for market basket

The figure above demonstrates that conclusions made from association rules are not always useful. Inexplicable results are those that can not be used for understanding of customer behavior because they are illogical and do not suggest further ways of action. These outcomes are also possible and should be excluded. In the given example, the most popular item sold on the day of opening of the new cosmetics store was washing powder. This shop did not find any ways to gain profit from the opening. Many questions arise. Why clients preferred washing powder? Was there

a huge discount on this product? Is it common reaction of customers? Were other marketing campaigns of cosmetics a failure? There are many unclear details related to the outcome, which can not be solved with the use of association rules and attempts to interpret this result. That is why results are inexplicable.

Trivial results are those that are obvious to every player in the market in this industry, such results also do not bring any extra value. Not only retailers and marketers, but also regular people are aware of the fact that smoked fish is a good snack with beer, and it is very common to buy these two products together. This knowledge is not unusual and does not bring any extra value for business.

Useful outputs are the goal of any analysis. In this case they should be informative and actionable. If customers buy wine and books on Saturday, it can be interpreted already creatively that people want to relax during the weekend reading interesting books and drinking wine. This finding can be a move to action, for example, to explore which segment of customers makes these purchases, based on that merchandisers and marketers can create new advertisements, etc.

Practical Example

Boztug (2017) in his research claims that iterative K-means clustering algorithm is the post appropriate approach for conducting market basket analysis. The procedure of analysis, offered by this researcher can be explained with following steps:

- Taking input data set, it is necessary to randomize sets of prototypes P by drawing K "seed points";
- Calculate distances between randomly chosen market basket vector and each prototype. The formulae for this calculations is an extension of the Jaccard coefficient, which is targeted to measure the "distance between a binary market basket vector and a real-valued prototype".

$$d(x_h, p_k) = 1 - \frac{(x_h, p_k)}{\|x_h\|^2 + \|p_k\|^2 - (x_h, p_k)},$$

  Where $d(xh, pk)$ represents the scalar results of vectors *xh* and *pk* respectively.
- Identify prototype with the minimal distance;
- Change it, according to the formula:

$$p_k^* := p_k^* + \alpha(\tau)(x_h - p_k^*),$$

  Where $\alpha(\tau)$ is a rate that is droningly declining with iteration time τ;

- Repeat the steps till the maximum number of iterations is reached.

Implementing this calculations in practice, the author got category choice probabilities, which were transformed to a visual model. This visualization is shown below (figure 3).

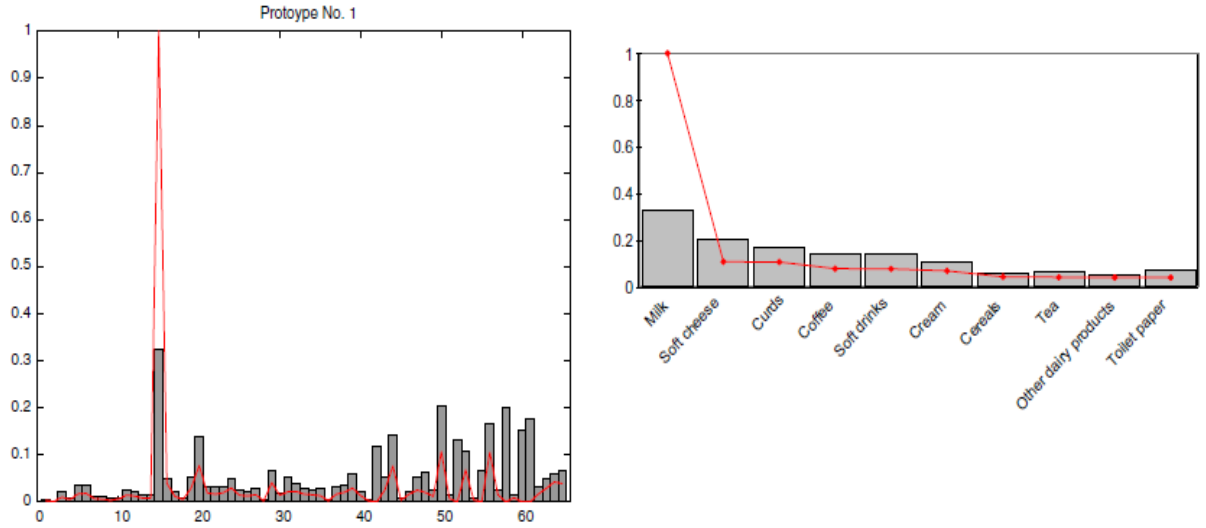

Fig. 3 Example of visualization of the market basket

This approach has a number disadvantages. The visual model itself is not the best example of models, which can be used effectively for further interpretation. According to the left diagram, the higher the bar, the higher unconditional purchase probabilities. This bar does not represent anyhow the market basket, it just shows that there are some potential baskets, as some of the bars are higher than other bars. The chart on the right side of the figure looks more informative. It can be interpreted that the potential basket could contain milk, soft cheese, curds, coffee, soft drinks, cream, cereals, tea, other dairy products and toilet paper. However, even this example demonstrates that there is cross-category purchases interdependencies between milk and soft cheese, but other purchase probabilities are even lower than 0,2. It means that approximately every 10^{th} customer will add to the shopping cart curds together with milk and soft cheese.

Another strong drawback is that this simple model is not applicable for Big Data. The exact number of transactions analyzed by the author is 69,736. This is a very small sample in terms of hypermarkets. Big stores can have this amount of transactions daily. It will be extremely time consuming to analyze market basket every day, this will be a senseless and expensive work because analysts will be paid for this amount of work conducted. The researcher Zerhari (2015) strongly criticizes the use of different types of cluster analysis as a tool for Big Data analysis because he believes that this type of algorithm is outdated and does not meet the requirements of capacity of processing of Big Data. This algorithm is implementable on smaller scales but not for

enormous volumes of data. Kurasova (2014) supports the idea of the author Zerhari that clustering is not appropriate for Big Data analysis. More developed technologies and experimental algorithms are required for Big Data.

Cluster algorithm is a widely spread statistical tool for segmentation. In this example, it was used, in order to test market basket analysis with some modifications of basic formulae proposed by Boztug (2017). However, this approach has disadvantages, that is why there is a space for improvement of this type of analysis. Even though different researchers have already criticized clustering as a method for manipulations with Big Data (in 2014, 2015), the author Boztug (2017) still used this algorithm as a base for market basket analysis and creation of further visualization. Visualized models were also not very representative in terms of data. Standard tools for visualization are not applicable for Big Data. The reasons for that were described in previous section.

1.4 Summary of theoretical part and justification of research gap

Literature about Big Data is popular, however, most of the articles interpret same information just using different words. At first sight, it might seam that the amount of articles and researches about Big Data is huge but the content is similar. A lot of researchers write about importance of Big Data in modern world, opportunities and threats. Enormous amount of data keeps increasing every second in different fields, it is getting harder to explore it and apply the information gained for the benefits of various organizations. Scientists, managers, retailers understand that future success and development depends on abilities to use Big Data wisely, handle it and extract value from it.

Data visualization is a good way to communicate information. This tool can be used in various fields for interaction with people with different background. Visualization models are frequently implemented to business goals. Visual models depict important data, show hidden interconnections that can be valuable for business but not obvious in other forms of information representation.

Retail chains face the problem of Big Data analysis, due to increasing data in the repositories about transactions and customers. However, this data should be explored for the benefits of the business because unexamined data keeps a lot of precious information. One of the dimensions for exploration is market basket analysis. This approach is not recently invented, based on simple association rules, but it is underestimated. The problem is that it is conducted with the use of common statistical tools and the main algorithm for market basket on practice is cluster analysis.

In order to identify market basket structures at advanced level using Big Data, more complex algorithms than cluster analysis should be implemented. This is a problem to be explored deeper. Every retailer faces the issue of increasing amount of transactions daily, which can not be left without analysis, due to huge potential losses of hidden valuable information in data. The topic explored is actual for the whole industry. It is a need for retailers to handle Big Data and understand customer behavior. However, new tools required to simplify the difficult task of data interpretation.

Based on the literature review and exploration of existing models, research goal can be formulated as follows: To develop a new model that will improve imperfections of the existing model and visualize effectively the outcomes.

Research gap: Existing models are not applicable for Big Data. As a result, there is a need for exploration of this field.

Chapter 2. Methodology of Data Visualization Techniques

2.1 Research design

For this research, both **qualitative** and **quantitative** methodologies are used. Quantitative research implies precise measurements. This methodology can be used to measure consumer behavior, opinions, attitudes or knowledge. As market basket analysis is one of the tools for understanding of customer behavior, quantitative research is appropriate method of research in this case [Cooper, Shindler, 2006]. Quantitative part was conducted with the use of SQL queries through programming tools. Main methodology was data mining in software Oracle Data Miner. Data mining techniques are related to quantitative type of research because there are hidden statistical formulae and functions in created program code for manipulations with data. In more details, some of the techniques will be described in this chapter.

In this business research, however, apart from quantitative methodologies, qualitative approach is also used. Qualitative research is an "array of interpretive techniques which seek to describe, decode, translate, and otherwise come to terms with the meaning, not the frequency, of certain more or less naturally occurring phenomena in the social world" [Cooper, Shindler, 2006]. Qualitative studies are a good option for researchers that will be involved in the process of data framing and interpreting. This is what is necessary in this research at the stage of data transformation and visualization. One of the main tasks of the researcher in this master thesis is to extract valuable association rules, interpret them. Creative interpretation by the researcher is also needed at the stage of data visualization, when the findings from visual models should be described.

As this work involves both quantitative and qualitative methodologies, this approach can be defined as mixed research design. In this research in-depth interviews are conducted simultaneously during the stage of data collection and analysis of this work. This means that the design of this research can be defined as a **partially integrated mixed method research design** [Saunders, Thornhill, 2016]. Apart from this, the research is based on the example of the researched company; this is also a **case study**. The name of the organization can not be disclosed in the paper because data used for this research is a trade secret. In the figure below it is shown, at what stages of the research qualitative, quantitative or both methods are used (Figure 4).

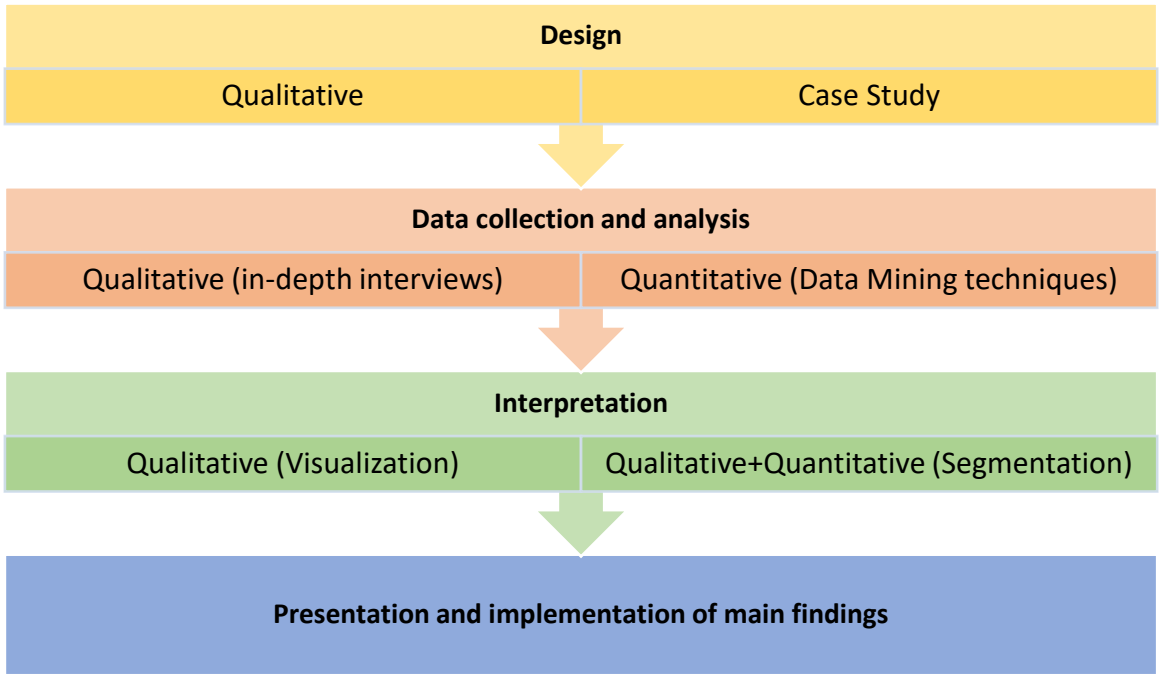

Fig. 4 Partially integrated mixed method research design

Most data sources used in this research are secondary. Studies that are made by others for their own purposes are assigned to the group of secondary data sources [Cooper, Schindler, 2006]. Secondary data can be extracted from internal sources and external sources. If the work is done for the company, or there is a case study devoted to particular companies, a lot of information by the organization is provided from internal sources. This can be annual reports from different departments, catalogues, brochures, etc. External resources are published data sources, such as industry statistics, books, articles, research reports [Ghauri, Gronhaug, 2005].

There are many advantages of using secondary data. As data is already collected in the organization and analyzed in the report it is of a high quality and can be used as the base for further research. Reliability of data is not the only plus of secondary data sources. Different books, articles can give a great overview about international researches because there is an access to literature of authors from different countries. Exploration of secondary data is also inexpensive because a lot of information is available for free; it is also time-saving, as it is relatively easy to access different data sources or they can be provided by the organization [Ghauri, Gronhaug, 2005].

Advantages are strong, however, some drawbacks still exist. For example, the amount of information available in different data sources is huge; the problem of choosing particular articles has to be solved because it is impossible to cover all the literature related to the topic. Moreover, some researchers might explore particular concepts but they have different objectives of the research and context. In this case, it is important to identify, whether these data sources are

appropriate as the base for the new research. Some problems with secondary data sources arise, however, high quality scientific researches can not be completed without using them.

Primary data also plays an important role in this research. This type of data is collected from the original source at hand. In this research, main way of primary data collection was in-depth interviewing. Asking people about the problem can give valuable information that can not be observed in the literature. The weakness of primary data collection could be unwillingness of the respondents to contribute to the research; however, this study is not the case. In-depth interviews are conducted with the people who were interested in the outcomes of the research; that is why their responses were professional and efficient, as that was beneficial for everyone.

2.2 Data Mining Methodology

Huge volumes of data are stored but hardly ever examined. This leads to the tendency that the world is becoming "data rich but knowledge poor". Unexplored data might be crucial for the company's prosperity or destruction, and organizations put a lot of effort into data analysis, in order to stay successful in the market, be innovative and keep competitive advantage. [Bramer, 2013]

Researched company in this master thesis is not an exception because there are many investments in the projects related to data mining and analysis. The goal is to find hidden valuable data that could be interpreted and used for the company's benefits.

What is data mining? Data mining is terms of business research is the "process of discovering knowledge from databases stored in in data marts of data warhouses" [Cooper, Schindler, 2006]. Data Mining can be also defined as a business process for exploration of huge amounts of data and identification of meaningful unobvious patterns and rules. [Berry, Linoff, 1997]

There are some imperfections in this methodology. Amount of data to be explored is usually enormous, and data mining is based on the data available, which is transformed and structured for the future analysis. This approach is widely spread, however, still imperfect because it is just an assumption that obtained results on limited data are applicable to the huge volume of unexamined data. To improve this, further data mining on unexplored data should be conducted. This process can last forever because the volume or data is increasing day by day, that is why a great number of assumptions are stated anyway and left without attention. For now there are no tools that give an opportunity to cover all volumes of data, however it is better to explore data and extract useful information from it than not to analyze it at all.

Exploring "raw" data, it is important not to forget about the goal of data mining, which is devoted to the search of meaningful patterns that have to be beneficial for business. The purpose is not about the process of mining itself, it is all about the value extracted from the data.

There are two basic approaches for data mining: hypothesis testing and knowledge discovery. Hypothesis testing is a so-called top-down approach because it starts with assumptions that will be approved or declined after data analysis. Knowledge discovery is different: it starts with the data; it is bottom-up approach. [Berry, Linoff, 1997]

Hypothesis testing

Hypothesis itself is a supposition or explanation made on the basis of limited evidence as a starting point for further investigation. [BusinessDictionary.com, 2018]

To understand the process of hypothesis testing, it is possible to divide for clarity this process into simple steps:

- Generation of ideas and formulation of hypothesis
- Determination of data needed for testing of the hypothesis stated
- Location of necessary data
- Transformation of data for future analysis
- Creation of models based on tested data
- Evaluation of the results
- Confirmation or rejection of the hypothesis based on the results obtained [Berry, Linoff, 1997]

This is a very simple explanation of hypothesis testing approach in data mining. Main difficulties of data mining lie in data structuring, transformation of it for effective analysis and interpretation of the results.

Knowledge discovery

Knowledge discovery can be defined as non-trivial extraction of implicit, previously unknown and potentially useful information from data. [Bramer, 2013]

Knowledge discovery used to be considered as data mining approach, however, in recent years, these two terms were distinguished. Knowledge discovery in databases (KDD) is an overall process of extracting valuable knowledge from data. Data mining is one of the steps in knowledge discovery process, which is based on the algorithms for retrieval of patterns from data. [Zanin, Boccaletti, 2016]

Knowledge discovery can be divided in the following stages:

- Data acquisition
- Data pre-processing
- Data transformation
- **Data Mining**
- Data post-processing (Interpretation/Evaluation) [Manco, Paturzo, 2016]

In this research KDD (Knowledge Discovery in Databases) approach is used. However, for further development of the outputs of this work and creation of visualizations, after data post-processing stage, two additional steps should be added: Data transformation (again) and representation of patterns. Data cleaning process is held after data acquisition, so this step also

should be added (in the second place). This approach step-by-step is represented in the figure below (figure 5).

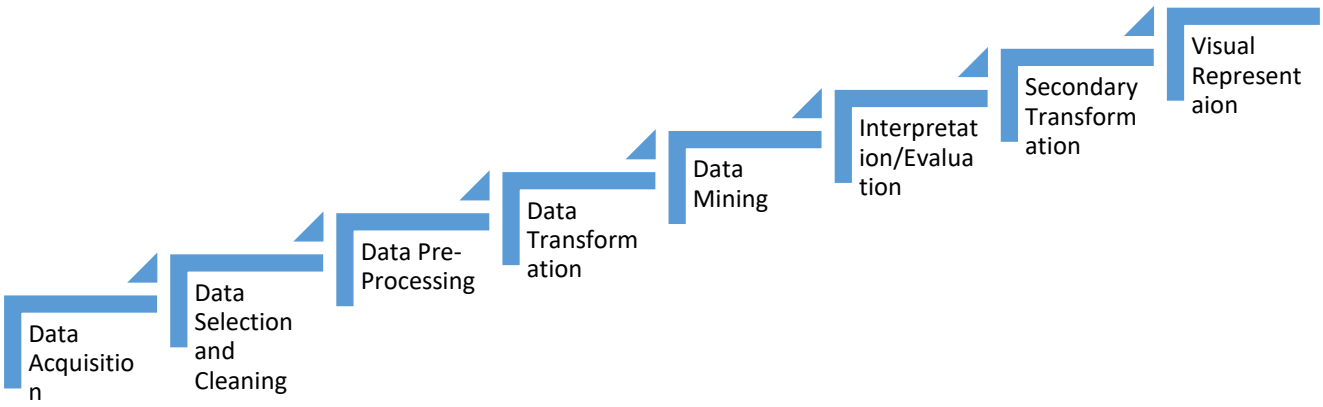

Fig. 5 Knowledge discovery process

Knowledge discovery process can not be successfully implemented without support of data mining system. In order to achieve better results in knowledge discovery, this system should meet following requirements:

- User-friendly graphical user interface that provides simple tools for analysis. Opportunity to work on basic tasks and steps helps to explore complicated data through visual models.
- For efficient data acquisition data mining system should be able to collect data from different data sources.
- Wide range of techniques for data preprocessing are crucial for such systems. The most time consuming and difficult task for data analysts is related to data selection, cleaning, preprocessing and transformation. To reduce time of these stages of analysis, huge variety of techniques for data preprocessing would be helpful.
- Apart from data preprocessing tools, data mining algorithms should be also developed.
- Data visualization tools for data interpretation, evaluation and representation.

- Expansibility of the system also plays an important role. Open architecture of the system would be a great advantage for adaptation of it to distinct tasks.
- Scale: for Big Data analysis, data mining system should be able to deal with mass-memory resident data.
- Wide range of data types should be recognized by the system. Techniques and tools for manipulations with different types of data should be developed. [Manco, Paturzo, 2016]

Researched company for Big Data analysis uses Oracle Data Miner GUI (graphical design interface).

Data Mining Functions

Overall, data mining functions can be divided into two main groups: supervised and unsupervised. Supervised functions can also be called directed. This type of functions require the specification of the target, in other words, known outcome, in order to make predictions about the value. The examples of directed data mining are:

- Classification: one of the options for grouping items and identification to which of discrete classes, the item may belong to;
- Regression: Estimation of relationship between values;
- Attribute importance: Highlights the key attributes for the predictive analysis;
- Anomaly detection: Search for outliers in a dataset.

Unsupervised functions do not require known outcome. These functions are used for identification of hidden interconnections in data.

Unsupervised data mining can be:

- Feature extraction: combination of the existing attributes and creation of new ones based on them;
- Association: Designed for market basket analysis;
- Clustering: Natural grouping of datasets [Taft, Stengard, 2005]

To meet the goals of this research, clustering and association functions are tested.

2.3 Oracle Data Mining

Oracle Data Miner GUI (graphical design interface)

Oracle Data Miner GUI meets all the requirements of a good data mining system listed in the previous section. It is a perfect tool to create predictive models, deploy various analytical methodologies, analyze huge amount of data, combine different formats and visualize the outputs. Oracle data Miner is also able to automate created algorithms, share and schedule them. These options simplify the work of analysts and give an opportunity to work with more complex tasks and data. Another advantage of Oracle Data Miner is generation of SQL scripts. This expands opportunities of analysis and automatization for the users.

Oracle Data Miner has a variety of useful built-in functions for data transformation and analysis. As it was discussed in previous section "Data Mining Methodology", for the goal of market basket formation and further segmentation of customers, at least two functions can be used. One option is cluster analysis, another one is association (based on the Apriori algorithm). The fragment of the interface of the built-in functions in Oracle Data Miner are represented in the figure below (Figure 6).

Fig. 6 Oracle Data Miner built-in functions

Clustering is a useful technique for finding natural grouping. A cluster itself is an assembly of similar data objects. In Oracle Data Miner, clusters are placed in the hierarchical tree type of model. The base for clustering in Oracle Data Miner is the algorithm: Enhanced K-Means algorithm.

Enhanced K-Means algorithm is a distance-based clustering algorithm, which relies on distance metric between the given points in the dataset. This metric measures the similarity between data points. Oracle Data Miner conducts enhanced K-Means algorithm in a following order:

- The hierarchical top down model is created with the use of binary splits;

- The algorithm designs each node of the tree step by step until the number of clusters customized by the analyst is reached;
- Data is distributed to clusters and the probabilistic scoring is provided;
- The algorithm returns the results for each of the obtained clusters, describes the hyperbox with the data assigned to the given cluster.

This approach to enhanced K-Means algorithm provides advanced results compared to classical K-Means. The interface is convenient to the user, so it is easy to run this algorithm.

Association is a perfect tool for market basket analysis. It is based on Apriori algorithm. This type of algorithm identifies the frequent individual items, and extends them to bigger item sets if they come out in the database frequently. This algorithm can be used as association rule determinant; that is why it is also applicable to market basket analysis.

Properties of association models can be calculated by Oracle Data Miner in two ways. Support of a rule can be used, which measures the frequency of items involved that occur in the rule. Another key property is confidence, which shows the conditional probability of how often the rule has been found to be true.

To understand how the built-in functions work in practice, it is necessary to test them. Main disadvantage of using standard functions is the inability of the intermediate control of data by the analyst. Data mining systems are developed enough for model creation, however, on a relatively small scale of data. In the case of huge retailers, data analysts have to deal with Big Data; and the role of analysts in data "cleaning" and transformation is crucial. Analysts can see some hidden correlations, filter out incomplete rows in data using program code, identify some "strange" results, which have to be double-checked. By using built-in functions, the risk of obtaining inadequate results because of "dirty" data is very high, that is why such results are not reliable for decision-making. This is the first reason for avoiding use of built-in functions for market basket analysis.

Another problem is a limited choice of parameters for running the algorithms. Some hidden correlations can be found only by using non-standard approach. The analyst is able to set some of the parameters, but all of them are standardized and simplified. This frames the outcomes in advance and makes the process of market basket analysis primitive and ineffective.

To illustrate this disadvantage, an example will be provided. Cluster analysis with the use of built-in functions in Oracle Data Miner was conducted. Data was not "cleaned" by the researcher in advance on purpose, in order to demonstrate the work of the algorithm. The result can be seen in the figure below (Figure 7).

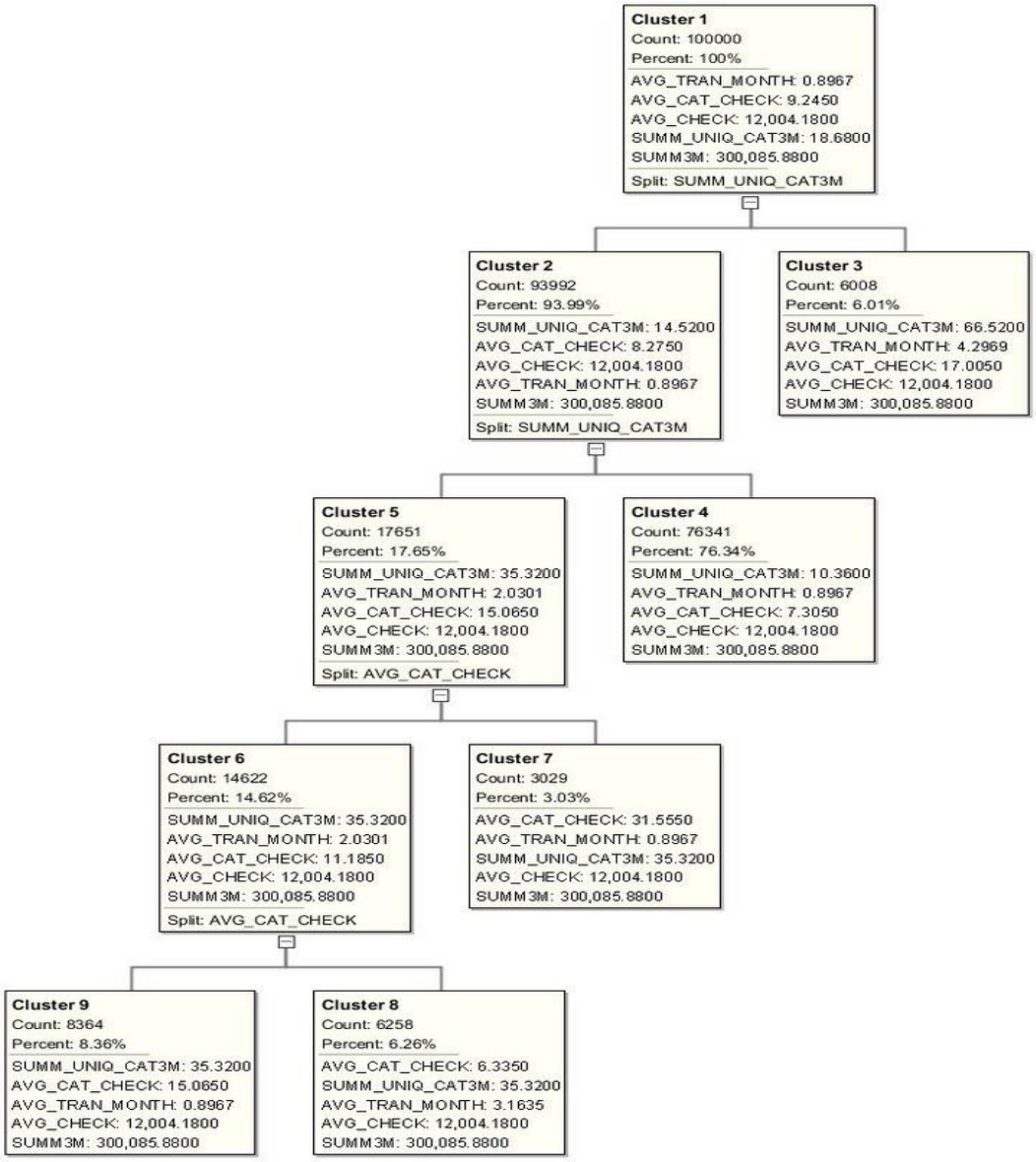

Fig. 7 Clustering in Oracle Data Miner

Comments on abbreviations:

- AVG_TRAN_MONTH – Average transactions made by the customer in hypermarket per month;
- AVG_CAT_CHECK – Average amount of categories stated in the receipt (check);
- AVG_CHECK – Average sum for purchased goods;
- SUMM_UNIQ_CAT3M – The sum of unique categories in the checks for three months;
- SUMM_3M – Sum for three months;

Imperfections of the results are obvious. One of the biggest drawbacks is that “SUMM3M” did not change, same problem with “AVG_CHECK”; these parameters were not adequately interpreted by the algorithm. This means that the data was “dirty” and the algorithm ignored this part without warning that there are problems with data. Apart from that, the algorithm does not exclude missing data automatically. As a result, it is not possible to estimate the quality of data and, based on that the quality of clusters obtained is also not measurable. Cluster analysis was run because one of the research questions of this thesis is devoted to segmentation. The goal was to show the problem of using cluster analysis as a tool for segmentation on Big Data.

Another built-in function to be tested for market basket analysis is “Association” button. As it was described earlier, this function is based on Apriori algorithm. To make the outputs easier for comprehension, the results were transported to MS Excel. A small fragment of the table obtained is presented in the table below (Figure 8) Bigger part of this table will be shown in the Appendix. Overall, this table consists of 3000 rows (and the results are already associations – “cross-selling pairs”).

	A	B	C	D	E	F	G	H	I
1	ID	Antecedent	Consequent	Lift	Confidence(%)	Support(%)	Item Count	Antecedent Support(%)	Consequent Support(%)
2	999	300	481	4.821	22.465	2.526	1	11.245	4.66
3	1602	558	554	5.427	38.379	3.423	1	8.92	7.072
4	1580	558	552	6.205	32.661	2.913	1	8.92	5.264
5	5297	791	779	4.77	12.643	1.126	1	8.906	2.651
6	5301	791	790	5.663	13.139	1.17	1	8.906	2.32
7	5299	791	785	6.457	13.122	1.169	1	8.906	2.032
8	5353	939	941	5.1	39.381	3.376	1	8.572	7.722
9	5085	570	573	5.561	42.748	3.574	1	8.362	7.686
10	1460	635	455	5.685	15.696	1.28	1	8.156	2.761
11	1311	454	402	6.543	19.582	1.527	1	7.796	2.993
12	5354	941	939	5.1	43.716	3.376	1	7.722	8.572
13	56	941	135	4.826	26.398	2.038	1	7.722	5.469
14	5086	573	570	5.561	46.504	3.574	1	7.686	8.362
15	1369	447	552	5.024	26.446	1.958	1	7.404	5.264
16	1373	447	555	4.798	16.86	1.248	1	7.404	3.514
17	700	108	881	5.507	35.012	2.549	1	7.281	6.358
18	710	108	944	5.338	16.16	1.177	1	7.281	3.027
19	1601	554	558	5.427	48.409	3.423	1	7.072	8.92
20	1578	554	552	7.346	38.667	2.734	1	7.072	5.264
21	2917	288	287	5.376	19.478	1.333	1	6.842	3.623
22	7191	193 AND 201	287	4.739	17.171	1.165	2	6.785	3.623
23	5216	690	681	5.713	32.712	2.176	1	6.652	5.726
24	9345	309 AND 968	969	5.027	57.627	3.831	2	6.648	11.464
25	1321	420	508	4.765	20.61	1.348	1	6.542	4.325
26	1329	420	555	5.289	18.587	1.216	1	6.542	3.514
27	701	881	108	5.507	40.093	2.549	1	6.358	7.281
28	9346	309 AND 969	968	4.878	61.328	3.831	2	6.247	12.571

Fig. 8 The fragment of the result of association built-in function transported to Excel

For example, analysing the first line of the table, it can be seen that the cross-selling pair of the segment number 300 and the segment number 481 was obtained. Number of categories are informative by themselves, it is needed to identify, what type of categories they are.

To find this out, another table with all the nomenclatures has to be used. Searching results for the segments 300 and 481 are stated in the figure below (Figure 9).

	A	B	C	D	E	F	G
1	Categorie's number	Order	Level	Number of purchases	Category	Segment	Market
13	300	12	Category	640753	Chocolate Bars	Chocolate Confectionary	Grocery
62	481	61	Segment	261355	Sweets and candies	Sugar Confectionary	Grocery

Fig. 9 Example of the search in the nomenclature

From the granted example, it can be concluded that one of the cross-selling pairs obtained is the association “Chocolate Bars –Sweets and candies”. These products are frequently purchased together. This approach is appropriate for identification of cross-selling pairs, however it is time-consuming to check each category of the goods in a separate table. Moreover, the process of creation of the whole market basket will be complicated based on the results obtained.

Both built-in functions are simple to use, however, relying on them only, it will be challenging to work on Big Data and identify not just cross-selling pairs, but also the whole market basket. That is why for the further research, more advanced tools with the use of SQL Queries are used.

2.4 In-depth and semi-structured interviews

In order to achieve higher results and avoid rediscover of already existing findings, in-depth interviews with senior analyst and manager of the department of researched company were conducted. It was necessary because specialists who work in the company have internal information that can not be provided or which is not fixed in existing documents of the organization.

Interviews can be divided into three groups: structured, semi-structured and unstructured. For this research unstructured interviews at different stages of knowledge discovery were chosen. The explanation for that is following: during the processes of data transformation, data mining and evaluation, informal conversations with the manager and senior analyst were an inherent part of the process. Firstly, all the manipulations with internal data of the company had to be approved. Secondly, the value of preliminary results obtained during knowledge discovery process were always discussed with the colleagues. Some of the outputs obtained could seam valuable to the researcher, however they could be obvious for representatives of the industry. Therefore, in-depth interviews helped to identify which ideas and what data is applicable for further research.

In-depth interviews have strong advantages. This approach allows the researcher to adopt flexible design. This means that informal interviewing allows to explore interesting aspects of the study, which can arise spontaneously. Some additional questions can be asked if necessary, and at different stages of the research it is a good opportunity to fulfill the gaps in intermediary results. Interaction is also a plus of interviews. It is possible to clarify details of some aspects, see the reaction of the respondents, check understanding of the questions and the answers.

Some drawbacks, however, should be taken in a consideration. Unstructured interview is not easy to keep within a frame. Some responses can be too broad; interpretation of them can be complex.

Getting back to this research, for in-depth interviews there was no predetermined list of questions; all of them were asked on an ad hoc basis during the process of the research face-to-face.

To illustrate this, real example from this research is provided. In more details, data mining process will be described in Chapter 3. In this section the example is given, in order to illustrate in-depth interviews only.

Analyzing the data, the researcher made an assumption to create “vegetarian” market basket. For this purpose, all the categories that contained any types of meat in the transactions had to be filtered out with the use of SQL Queries. Another assumption of the researcher was that this filter was not enough to form the “vegetarian” market basket. Informal discussion helped to answer

a lot of questions, stimulated brainstorming and was the base of the decision about further interpretations with the data, taking in a consideration opinion of three people about the issue, regarding the "vegetarian" basket. As the questions of the in-depth interview are not structured and not formulated in advance, approximate formulation or the questions discussed is as follows:

- Should the filtering out of meat categories be the only basis for "vegetarian" basket?
- Creating another filter using SQL Queries, what categories should be included? Fresh Fruit, Vegetables, cereals, porridges (this was the assumption of the researcher)? What about frozen vegetables/fruit?
- Creating filters "vegetarian" basket, are there many categories in the second filter which could make the segment of this basket insignificant in the amount of all transactions analyzed?

These are just examples of questions stated during the data mining process. To solve some of the issues during the analysis, more than 20-25 informal question could be asked, in order to continue further exploration of data. Examples of interviews are provided in the Appendix.

Unstructured interviews played an important role in this study, however this type of interviews was not the only one involved in the research. Semi-structured interview was conducted, in order to identify the managerial implications of the completed work. Semi-structured interviews have similar pros and cons with unstructured interviews. The only difference is that semi-structured interviews are semi-flexible. In more details, this interview will be described in section "3.3 Managerial implementation of the results".

To illustrate the types of interviews at different stages of the research, the following figure is provided (Figure 10)

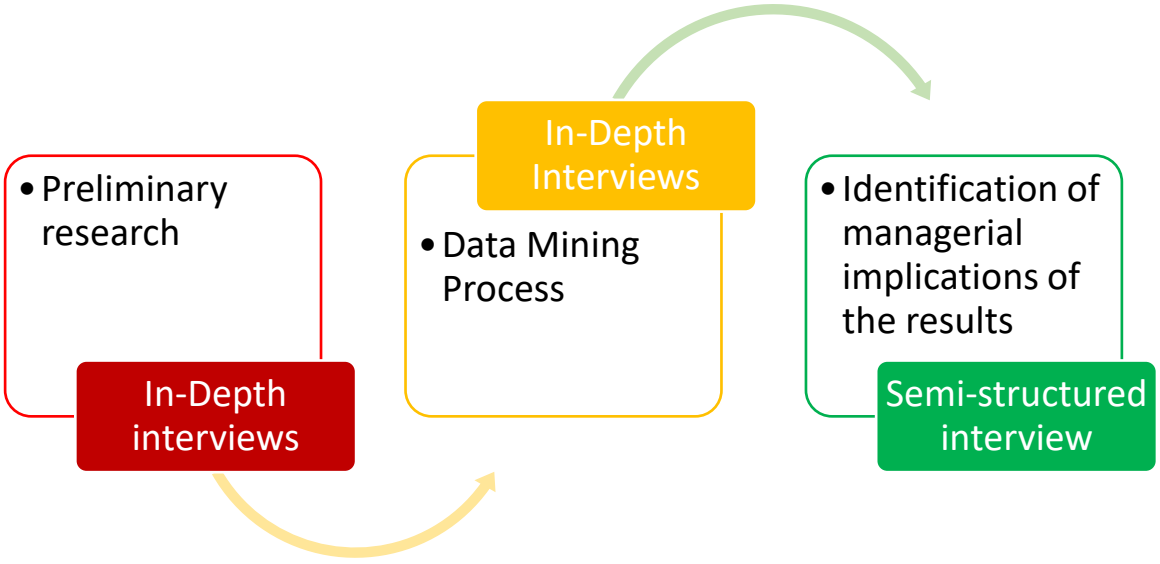

Fig. 10 Types of interviews at different stages of the results

Interviews play an important role at different stages of the research. The amount of valuable information obtained from the primary data and the openness of the respondents (senior analyst and manager of the department) helped to explore deeply the topic and achieve high results.

2.5 Summary of chapter 2

Main methodologies of this research are data mining (particularly, Oracle data Mining) and interviews (of two types: in-depth and semi-structured). Knowledge discovery was selected a data mining approach because it is meets the goal of the research. Knowledge discovery is a non-trivial extraction of previously unknown but potentially useful information from the huge amount of data, and this is exactly the case. Knowledge discovery process involve several stages: Data acquisition, Data pre-processing, Data transformation, Data Mining, Data post-processing. As the research will not finish on data mining process (visualization models will be created afterwards), two extra stages were added to knowledge discovery process: additional data transformation and representation of the patterns obtained.

Oracle Data Mining will be the main part of empirical part of the research. Oracle Data Miner as a mining system offers different built-in functions. Two of them: association and clustering are appropriate for market basket analysis and segmentation. Both of them were tested, and the results were not as great as expected. Clustering as a tool for natural grouping based on Enhanced K_Means algorithm ignores imperfections in the "raw" data, and the analyst in the output can not identify these mistakes. Moreover, the parameters selected as a base for clustering are limited, which is not efficient for search in data of some unusual clusters. Association as a built-in function was also not the perfect option. The cross-selling pairs were identified, however, creation of the whole market basket based on the outputs is challenging and time-consuming. Because of these disadvantages, it was concluded that use of built-in function is not reliable for Big data analysis without intervention of the data analyst. That is why it was decided to conduct the analysis with the use of SQL Queries in Oracle Data Mining.

Primary data is also crucial for this research. Interviews with the senior analyst and manager of the department of the researched company are supporting part of each stage of the research. In-depth interviews are conducted at the stage of preliminary research and data mining process. Semi-structured interview are concluding part of the research and tell about the managerial implications of market basket visualization.

Chapter 3. Market Basket Visualization

3.1 Preliminary data transformation

Efficient data mining process requires data cleaning and preparation for further manipulations.

In order to get efficient associations, which could be adequately represented in visual models, preliminary data transformation was conducted at the level of segments and categories before any procedures in Oracle Data Miner or Oracle Data Visualizer were done. Identifying of cross-selling categories at the level of segments only was not the right option because that would be an enlarged level of association creation and, as a result, outputs would have been uninformative. Decision of finding combinations of correlations at the level of categories was also declined because overall, there were 683 categories at the beginning. Identification of cross-selling relationships between big amount of categories will lead to complex visualization models that will be so overloaded, that correlations between categories wouldn't be visible. Apart from that, during the process of data mining, correlations between 683 categories (EACH-WITH-EACH) will have relatively small share in data, which means that some combinations might seam insignificant, however, they might have a great potential for analysis if they were explored at the smaller scale.

The solution is to create a balanced list of categories and segments together. Where necessary, enlarge the categories and unite them with the particular segment, if not – leave the category as a separate important item. The decision about merger of categories is based on assumptions of the researcher and in-depth interviews.

The process of elimination of categories can be illustrated with the example. It is provided for different cheese segments and categories. Some details should be clarified first. Cheese with proprietary packaging of the retail chain is cheaper than cheese packed at the dairy factory. Moreover, it is not clear for the customer who is the producer of the cheese packed and sliced at the hypermarket because only the type of cheese and data of packaging is stated. On the one hand, this makes with proprietary packaging of the hypermarket less trustworthy for the consumer. On the other hand, this type of cheese has lower price, and this is still same product.

Considering previous explanations, after in-depth interviews during the process of reorganization of the list of segments and categories, it was decided not to create one general segment of cheese, but to create two different segments: "Cheese and milk products (packaging of the retailer)" and "cheese with dairy factory packaging". The results of united segments are shown in the figures below (Figure 11 and Figure 12).

Nine different categories of cheese are distributed to the same segment. It is the matter of taste of customers, what type of cheese they prefer: cream cheese or Parmesan, etc. The most

important thing is that these types of cheese are produced at factories, and every package has a recognizable brand for clients compared to the cheese with retailer's packaging. The goal of the research is not about understanding of varieties of cheese mostly consumed by customers; it is devoted to market basket analysis and segmentation in general. That is why based on the in-depth interviews it was decided to create only two enlarged segments, not more than that.

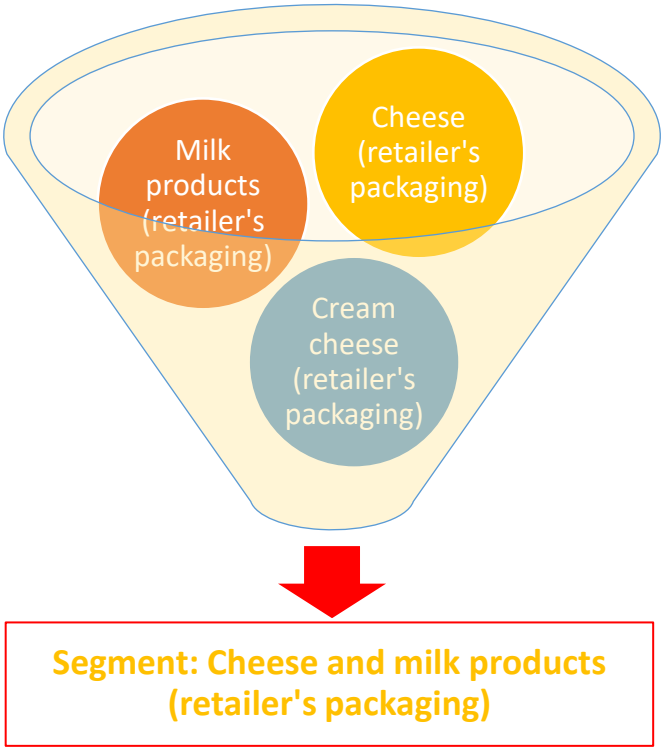

Fig. 11 Reorganization of the segments and categories

Cheese with dairy factory packaging	Young cheese
	Cream cheese
	Chopped cheese
	Hard cheese
	Fat-free cheese
	Goat's cheese and cheep cheese
	Premium cheese and fondue
	Cooled packed cheese
	Cooled packed cheese for degustation

Fig. 12 Reorganization of the segments and categories

Previous example demonstrates the principle of reorganization of segments and categories of products for further analysis. Previously, there were 683 categories; after transformation, only 157 were left. This reorganization of segments and categories of products is necessary for the final result of this research: creation of clear visualization models.

3.2 Experimental procedures of SQL Queries

Initial data was taken from the portal and transferred to Oracle Data Miner. The level of aggregation of input data was two months. In other words, transactions analyzed simultaneously were taken for the period of two months at two hypermarkets at the same time. Overall, market basket analysis was conducted for the period 01.01.2017 – 31.12.2017. Intervals of two months were chosen because the goal of the research is devoted not only to the client segmentation and identification of various market baskets, but also to analysis of hidden trends in data. Shorter periods could depict seasonal specific trends in customer's behavior. Another reason for the choice of the two-month interval for analysis was time saving. If data about transactions for the whole year in two hypermarkets was transferred to Oracle Data mining, the program would have needed days to finalize one simple procedure. Even with data for the two-months period there were issues sometimes because one procedure with data could last for several hours. This is a good demonstration of how "big" Big Data is and how many difficulties are there on the way to the exploration of enormous amount of data.

Market basket formation was based on assumptions created by the researcher. These assumptions were approved/disapproved after in-depth interviews with the manager of the department and the senior analyst. Later on, these assumptions were checked on data. If the market basket obtained had a significant share among all the transactions analyzed for the given period, then it was checked some more times. If this market basket was significant, this basket was formed and used for further analysis.

Overall, 20 different assumptions were checked. In more details, all the assumptions are described in the section "Interpretation of the results obtained".

The turn-based steps of the analysis can be described as follows:

- Uploading data from the portal to Oracle Data Miner;
- Data transformation for effective analysis (so-called data cleaning);
- Formulation of assumptions about potential market baskets;
- Creation of filters with the use of SQL Queries for application of assumptions stated;
- Identification of the significance of the outputs obtained (decision about further development of the assumption);
- Transformation of data using new united categories and segments;
- Recheck of data on a sample;
- Data transformation for exportation of it to Oracle Data Visualization;
- Creation of Visual models of the type Network.

Despite of the fact that the analysis was conducted for relatively short periods (two months), some manipulations with data could take a couple of hours for a single operation.

After data uploading from the portal to Oracle Data Miner, all the data is represented in the table with millions of rows. All the manipulations were made in the "workflow" in Oracle Data Miner. To illustrate step by step process, the workflow from Oracle Data Miner is shown below (Figure 13). The example provided is market basket analysis of "Vegetarian" basket (in more details it is described in section "3.3 Market Basket Visualization Results".

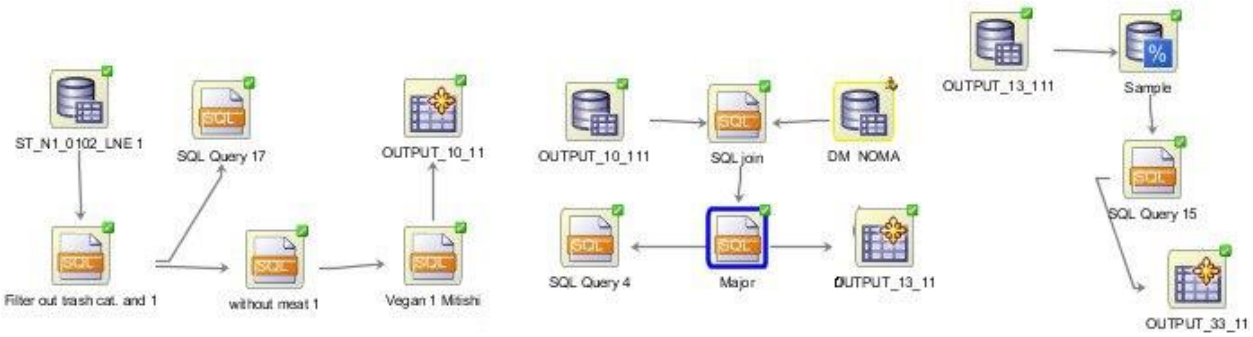

Fig. 13 Workflow in Oracle Data Miner. Example of "Vegetarian" market basket

From this figure 10, it can be seen that mostly there are SQL Queries, which means that most of the manipulations during data mining process were made with the use of SQL program code. Other icons "ST_N1_0102_LNE1", "OUTPUT_10_11", "DM NOMA", "OUTPUT_13_111" represent tables with data from the database. Icons "OUTPUT_13_11" and "OUTPUT_33_11" also contain tables, however the size of these tables is compressed.

As it was mentioned earlier, data was uploaded from the Portal to the data mining system; the first table obtained is "ST_N1_0102_LNE 1".

After that the process of data cleaning was launched. First of all, it was necessary to identify in the table of data all incomplete rows, which are defined as "dirty" data. Another important consideration was related to filtering out with the use of SQL Queries all "trash" categories, for example, those products that are not represented in the hypermarket in assortment, however, stated in the nomenclature. Data about transactions at cash desk for employees also had to be excluded from analysis because this group of customers purchased lunches at the particular cash desk for different prices. The challenge was to collect the information about the number of cash desks for employees and filter out this data from the table. The fragment of one of the SQL Queries is shown in the figure 14.

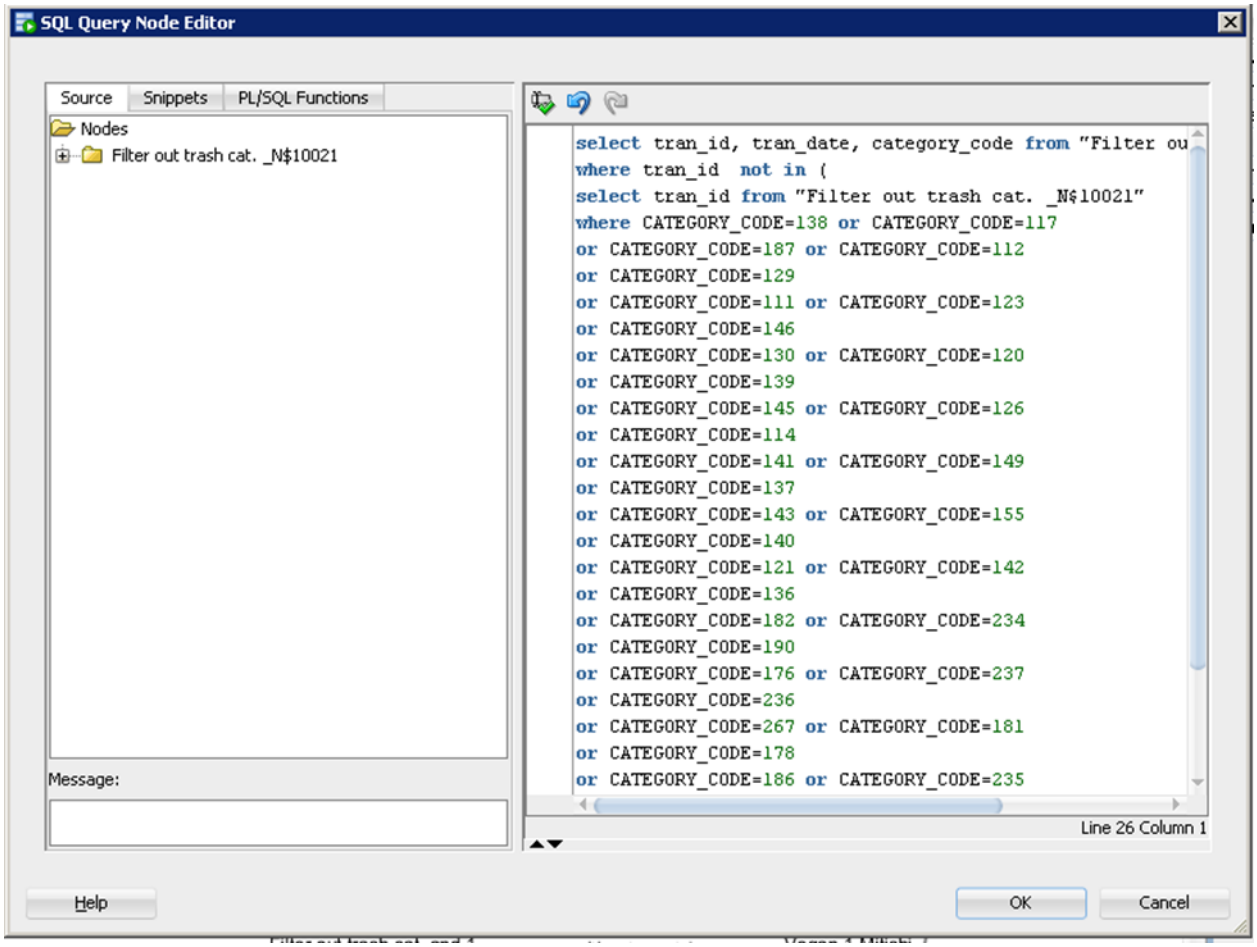

Fig. 14 Fragment of SQL Query for filtering out "trash" categories

Next step is reflection of the assumptions about potential market basket. For this purpose, different categories and segments were filtered out from data based on assumptions or, vice versa, added if necessary. SQL Queries "Without meat 1" and "Vegan 1 Mitishi" are these filters. First filter excludes all the categories and segments that contain meat products of different types in all departments of the hypermarket. The fragment of the program code can be found in Appendix. Second filter includes some of the categories of products that are expected to be in the "Vegetarian Basket". Each of the categories and segments has an assigned number in the existing database of the researched company. Detailed description of all the market baskets obtained is given in section "3.3 Market Basket Visualization results".

Following step "OUTPUT_10_11" is made, in order to increase the speed of data processing. Preliminary tables have to be uploaded at different stages of data mining, otherwise the process will start from the very first step in the chain and, as a result, this manipulation will be time-consuming. The issue of Big Data transformation and timing of analysis is always actual.

Previous table was uploaded as a separate database, and the analysis continued. Preliminary results of analysis (with filtered categories) have to be combined with the new database, which

contains new categories and segments, which were created at the stage of preparation for data mining. Detailed description of merger of segments and categories was given in the section “3.1 Preliminary Data Transformation”. This stage of data mining slowly overflows to the stage of data transformation for further visualization.

“SQL Query 4” after creation of the new database with edited categories and segments was conducted, in order to count intermediate amount of transactions. That is necessary for estimation of amount of data before it is visualized. If the number of consumers’ checks that pass to the created market basket has the share in total amount of transactions that exceeds 5%, the assumption about the potential market basket (in this case “Vegetarian Basket”) is estimated as significant, and the analysis of the outcomes keeps developing.

Again, the chain of SQL Queries becomes too long, it slows down data processing, that is why a separate table has to be uploaded.

All the stages of data mining process were held on the entire volume of data. Data Visualization, however, implies reduction of the amount of data, due to the limitations of Oracle Data Visualizer. After that random sample of transactions was generated, which includes 600 000 rows in the table.

Before final output table was created, one more operation with data had been carried out. The type of visualization models for market basket was expected to be “Network” because this sort of visualization could clearly depict the associations. Data had to be transformed in a way the system for visualization could recognize it, that is why “SQL Query 15” was created. This SQL Query formed a new column in the database, which contains only numeral “1” in each of the rows. This simple transformation is necessary for the visualizer, the algorithm of which will count the frequency of the cross-selling pairs, interpret these associations and depict them in the visualization model.

After all these procedures, if the assumption is approved based on the share of the potential market basket, the intermediary output of the customers’ basket can be viewed through data in the table. For example, the following figure (figure 15) shows the most frequent cross-selling pairs of products and the number of times these associations are met in the given sample. The figure represents information in Russian language, as the researched company provided all the data in Russian. All the abbreviations of the category’s names are also original (taken from the company’s data).

CAT1CAT2	COUNT_LINK
ФРУКТЫ ПОСТОЯННОГО АССОРТ<--->ПАСТ.МОЛОКО,К/МОЛ ПР-ТЫ	26,235
ФРУКТЫ ПОСТОЯННОГО АССОРТ<--->РАЗВЕС - БАКАЛЕЯ ЗАКУСКИ	25,095
ФРУКТЫ ПОСТОЯННОГО АССОРТ<--->ОВОЩИ ТЕПЛИЧНЫЕ	22,440
ФРУКТЫ ПОСТОЯННОГО АССОРТ<--->ДЕСЕРТЫ ОХЛАЖДЕННЫЕ	21,730
ФРУКТЫ ПОСТОЯННОГО АССОРТ<--->Белые хлебы об.	20,284
об.Прочие бытовые принадл. и средства<--->ФРУКТЫ ПОСТОЯННОГО АССОРТ	19,219
об.Прочие бытовые принадл. и средства<--->об. Средства для стирки	18,606
РАЗВЕС - БАКАЛЕЯ ЗАКУСКИ<--->ПАСТ.МОЛОКО,К/МОЛ ПР-ТЫ	16,560
ПАСТ.МОЛОКО,К/МОЛ ПР-ТЫ<--->ДЕСЕРТЫ ОХЛАЖДЕННЫЕ	15,977
РАЗВЕС - БАКАЛЕЯ ЗАКУСКИ<--->Белые хлебы об.	15,936
об.Прочие бытовые принадл. и средства<--->РАЗВЕС - БАКАЛЕЯ ЗАКУСКИ	14,879
ФРУКТЫ ПОСТОЯННОГО АССОРТ<--->ОВОЩИ ОТКРЫТОГО ГРУНТА	14,444
об. Злак. и диетич. батончики<--->ФРУКТЫ ПОСТОЯННОГО АССОРТ	14,389
ШОКОЛАДНЫЕ ПЛИТКИ<--->ФРУКТЫ ПОСТОЯННОГО АССОРТ	14,229
ПАСТ.МОЛОКО,К/МОЛ ПР-ТЫ<--->Белые хлебы об.	13,822
ФРУКТЫ ПОСТОЯННОГО АССОРТ<--->ПРОМ-Е ХЛЕБОБУЛ-Е ИЗДЕЛИЯ	13,576
РАЗВЕС - БАКАЛЕЯ ЗАКУСКИ<--->ДЕСЕРТЫ ОХЛАЖДЕННЫЕ	13,283
ОВОЩИ ТЕПЛИЧНЫЕ<--->ЗЕЛЕНЫЕ КУЛЬТУРЫ	13,282
об. Конфеты и леденцы<--->ФРУКТЫ ПОСТОЯННОГО АССОРТ	13,168

Fig. 15 Most frequent cross-selling pairs of products

From the Figure 12 an interesting observation can be made: in top five cross-selling pairs of the products, one of the categories in each of them are “fruit of permanent assortment”. Based on these results, the potential market basket can be distinguished, which contains of categories and segments “fruit of permanent assortment, pasteurized milk and milk products, white bread, grocery – customers’ package by weight, cooled desserts and greenhouse vegetables.

Translation of the fragment of the output from Oracle Data Miner is represented below (Table 1).

Category 1	Category 2	COUNT_LINK
Fruit of permanent assortment	Pasteurized milk and milk products	26 235
Fruit of permanent assortment	Grocery – customers’ package by weight	25 095
Fruit of permanent assortment	Greenhouse vegetables	22 440
Fruit of permanent assortment	Cooled desserts	21 730

Fruit of permanent assortment	White bread	20 284
Other household accessories	Fruit of permanent assortment	19 219
Other household accessories	Washing powder	18 606
Grocery – customers' package by weight	Pasteurized milk and milk products	16 560
Pasteurized milk and milk products	Cooled desserts	15 977
Grocery – customers' package by weight	White bread	15 936
Other household accessories	Grocery – customers' package by weight	14 879
Fruit of permanent assortment	Field vegetables	14 444
Granola bars	Fruit of permanent assortment	14 389
Chocolate bars	Fruit of permanent assortment	14 229
Pasteurized milk and milk products	White bread	13 822
Fruit of permanent assortment	Bakery products	13 576
Grocery – customers' package by weight	Cooled desserts	13283
Greenhouse vegetables	Leafy vegetables	13 282
Sweets and lollipops	Fruit of permanent assortment	13 168

Table 1. Translation of the outcome from Oracle Data Miner

Data mining techniques built on SQL Queries lead to impressive results, which could not have been reached with the use of standard functions of the data mining system. This experiment showed that creative approach that involves formulation of assumptions can be more effective than analysis with the use of built-in functions. Oracle Data Miner is a powerful tool for manipulations with data, however, to achieve higher results in Big Data analysis, it is necessary to expand the frames and use also SQL Queries.

3.3 Market Basket Visualization Results

The final table obtained after data mining process was uploaded to Oracle data Visualization. At first sight, the result of visualization seams to be uninformative and complicated. The output is shown in the Figure 16.

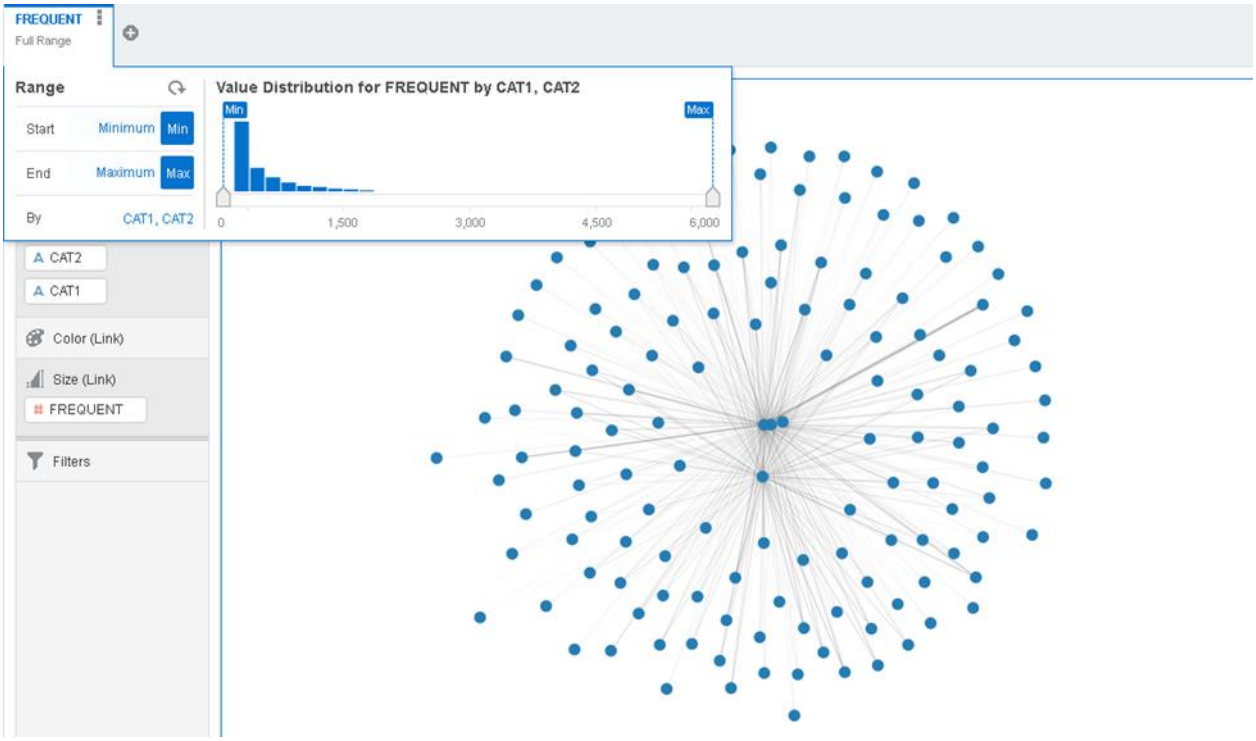

Fig. 16 Preliminary visualization of market basket

However, customizing the frequency (in the left upper corner of the interface), it is possible to adjust the given visualization model, and then the results get informative. In the figure below, one of the obtained market baskets is represented (Figure 17). The screenshot from Oracle Data Visualization is provided (original one from the program), that is why the output is in Russian language (company's data). Translation is shown in the figure 18.

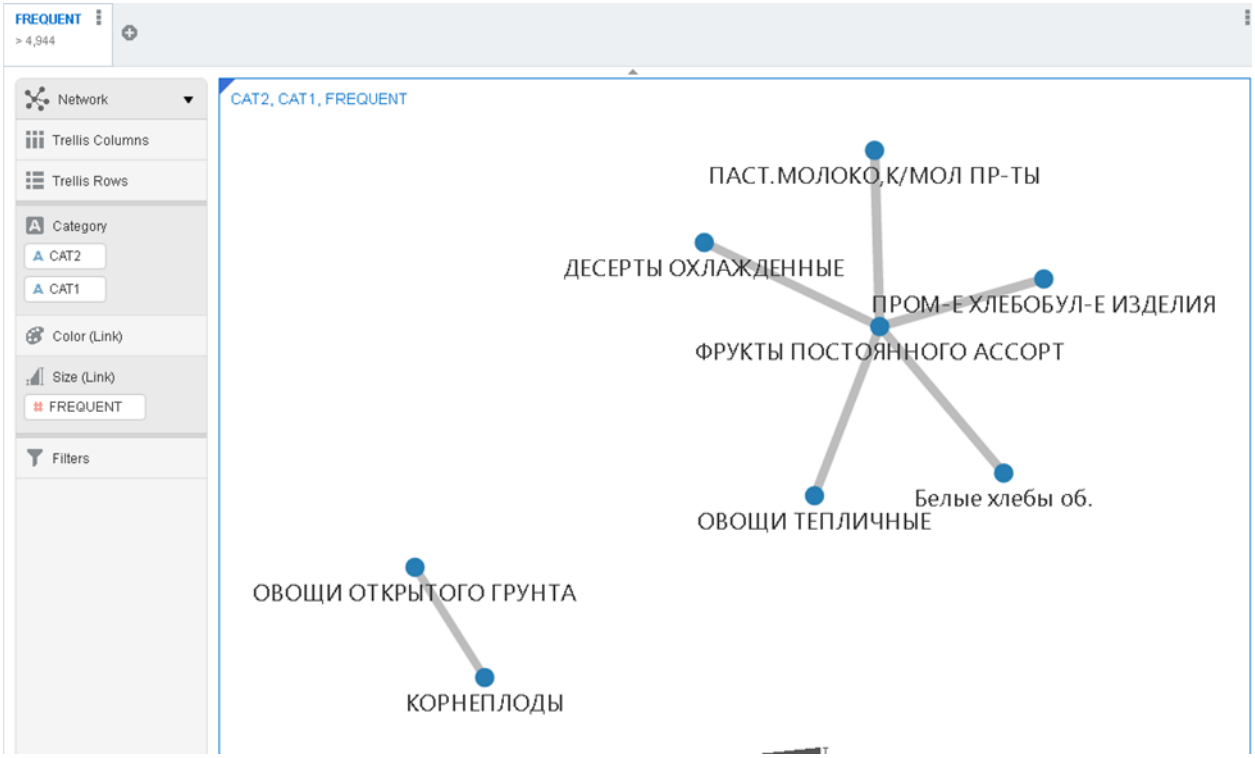

Fig. 17 Example of market basket visualization

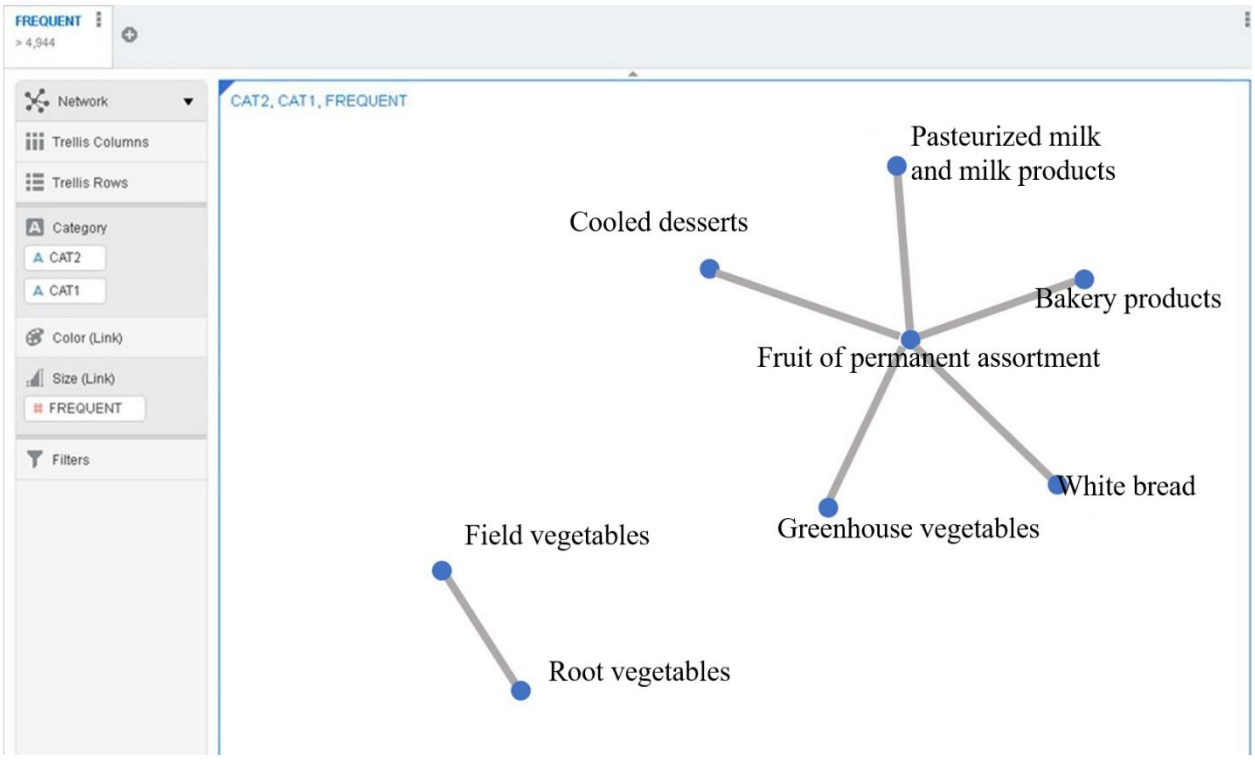

Fig. 18 Example of market basket visualization in English

Visualization of the market basket provides the analyst with a clear picture of the results. Compared to built-in data mining functions, which lead to meaningful results finding the cross-selling pairs, but they did not give the output of the whole customers' basket. Data mining techniques with the use of SQL Queries showed more outstanding results than built-in functions, and even the market basket that contained several goods was obtained, however, visualization

model helps to see a broader picture of the outcome. In the example of the visualized customers' basket provided, a bigger basket with six products in it can be identified. Apart from that, visualization helps to see important cross-selling pairs, which were not on the "top" of the table (data mining output) in a particular basket, but they are meaningful as a separate pair of goods or a small basket. To demonstrate ability of the visualization models to illustrate "small" but very significant baskets, the example is given (Figure 19). Translation is provided in Figure 20.

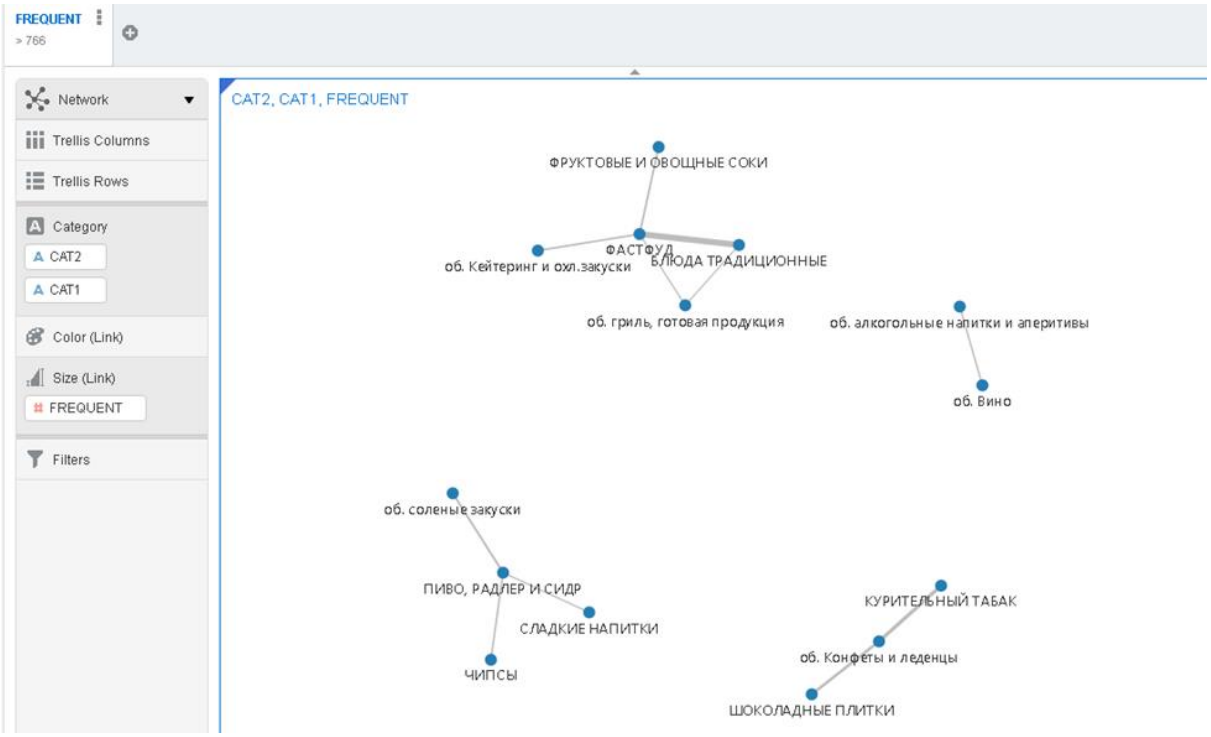

Fig. 19 Example of Visualization of the "small" baskets

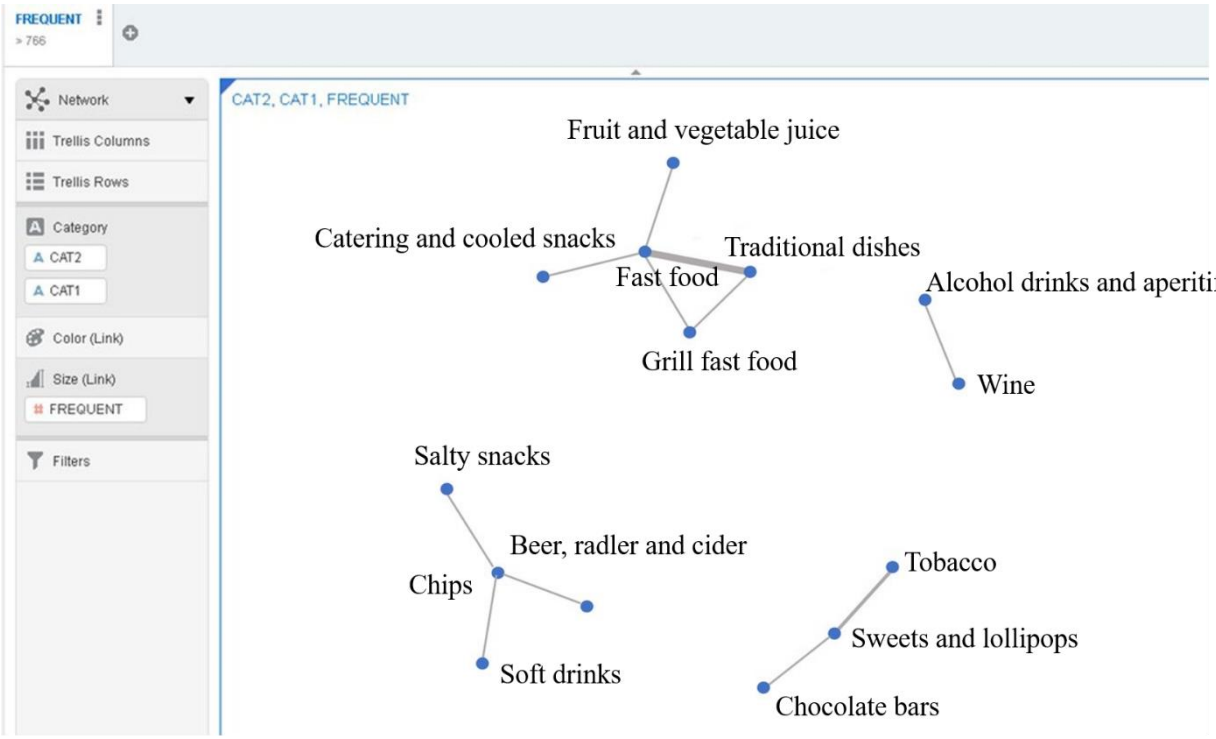

Fig. 20 Example of Visualization of the "small" baskets English version

Twenty assumptions about types of market baskets were formulated, and only seven of them were approved after data mining and following visualization. Filters that were used for SQL Queries, in order to exclude or include some segments of the products; assumptions, intersection of categories in checks, share of the particular type of customers' basket in all the transactions for the given period and the list of goods of the basket obtained will be clarified in the following table (Table 2).

Name of the market basket	Assumptions	Filters	Intersection with other baskets	Share (%)	Top products in the basket
Vegetarian Basket	Customers of this segment do not buy any type of meat products. However, they tend to eat healthy, that is why some healthy goods are purchased	Filter out all the segments/categories, which contain meat products, canned goods, poultry meat. Include only those transactions that contain at least one of the segments/categories: fruit, vegetables, frozen vegetables, cereals and groats. Consider only those transactions, which meet both requirements	Intersections are possible with all the baskets but "Basket for meat lovers" and "Non-food basket".	19%	1) Fruit of permanent assortment 2) White bread 3) Greenhouse vegetables 4) Cooled desserts 5) Grocery – customer's package by weight 6) Vegetables (roots) 7) Seasonal fruit 8) Frozen berries 9) Pasteurized milk and milk products 10) Whole wheat bread

Basket for Meat lovers	Fresh or frozen meat and fish are key components of this basket	Transactions that include at least one type of meat/fish category should be chosen	Intersections are possible with "Alcohol Basket", "French Basket"	24%	1) Sausages 2) Cheese with dairy factory packaging 3) Ham 4) Cooled chicken 5) Chopped meat 6) Eggs 7) Meat delicacies 8) Butter and margarine 9) Cooled turkey 10) Sauce for meat dishes
Non-food Basket	This basket includes different household goods and does not contain any food products	All "food" categories are filtered out	No intersections with other market baskets	10%	1) Toothpastes and toothbrushes 2) Washing powder and other laundry detergent 3) Toilet paper 4) Dish soap 5) Doormats, tissues, sponges for dishes

					6) Household chemicals for toilets and bathrooms 7) Soap and shower gel 8) Screenwash fluids 9) Shoe polish 10) Towels for the kitchen
French Basket	This is a small basket that obligatory has any type of wine in it	All checks that contain wine are selected	Intersection with other market baskets is possible but "Non-food basket"	7%	1) Red wine 2) White wine 3) Wine drink 4) Cheese with dairy factory packaging 5) Meat delicacies 6) Ham 7) Assorted seafood 8) White bread 9) Cooled desserts 10) Water with gas
Parent's Basket	This basket contains goods for	All transactions with segments that consist of goods for kids	Intersections are possible with "Vegetarian	17%	1) Diapers 2) Goods for hygiene of

	newborn babies	under 3 years old, books for children, toys are selected	Basket" and "Basket for meat lovers"		newborn babies 3) Toys 4) Washing powder and other laundry detergent 5) Baby food
Smoker's Basket	This basket obligatory incudes cigarettes and tobacco categories	Selected all transactions with the segment "Cigarettes and tobacco"; Filtered out top 10 goods from the previous baskets stated in this table	Intersections are possible with "French basket" and "Alcohol basket"	9%	1) Cigarettes and tobacco 2) Sweets and candies 3) Bubble gum 4) Carbonated drinks 5) Water
Alcohol basket	Main goods in this basket are drinks with alcohol	Filtered out top 10 goods from the previous baskets stated in this table but "French basket"; All transactions with wine, vodka and aperitifs	Intersections are possible only with a "French basket"	2%	1) Alcohol drinks and aperitifs 2) Wine 3) Vodka

Table 2. Market baskets obtained

Overall, these seven baskets represent 88% of the transactions. Some intersections are possible (and they are described in the table 1), however, this output is not only about market basket determination. Distribution of 88% of the transactions of customers can be considered also as segmentation through market basket analysis. Compared to other data mining methodologies with the use of built-in functions and the market basket analysis conducted by Boztug (2017) in

his experimental research (it is described in chapter 1 “Market Basket Analysis” section), creative approach to market basket analysis with the use of SQL Queries based on assumptions helped to identify effectively different types of market baskets. Apart from that, extra valuable output was obtained – this analysis showed an interesting segmentation of customers.

3.4 Managerial implementation of the results

To understand possible managerial implications of the results, semi-structured interview was conducted with the manager of the department of innovation of one of the biggest retail chains (as the research was based on the data of the researched company, the name of the organization can not be disclosed).

Five main questions arose during the interview. They are stated in the figure below (Figure 21)

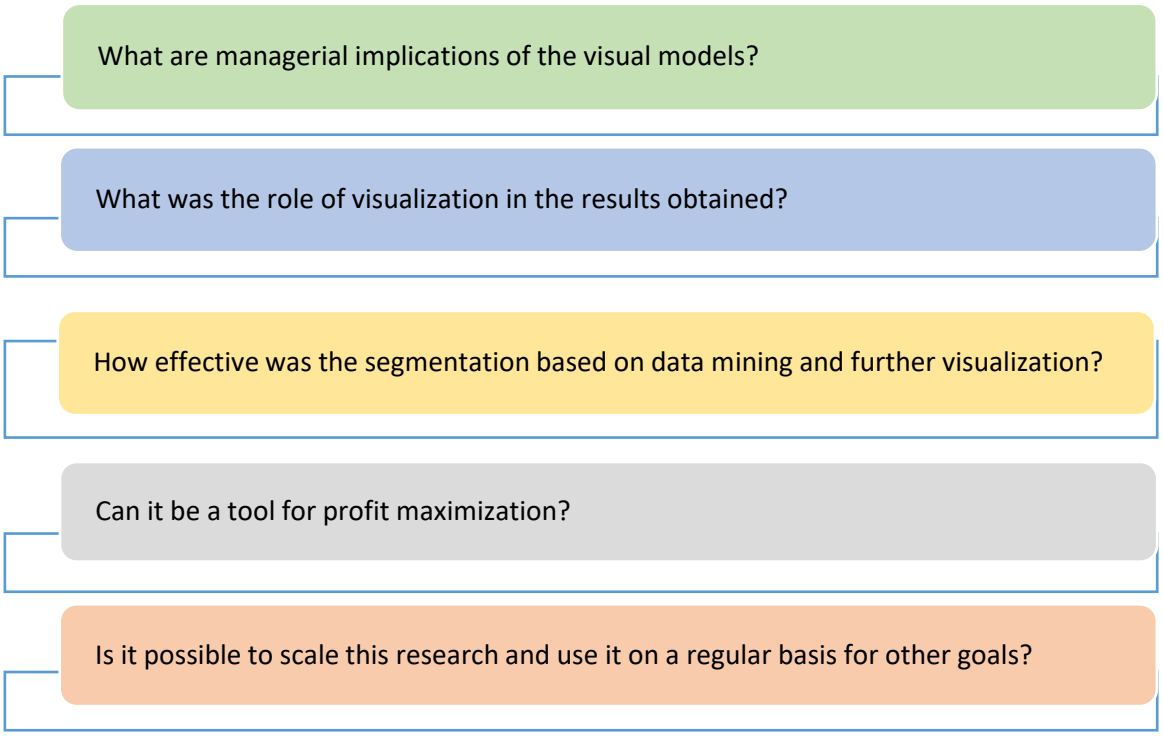

Fig. 21 Key questions of the semi-structured interview

The direct quotation of the open-added response is following: “For big companies it is crucial to demonstrate reports based on Big Data and represent them as different relevant excerpts (about location, format, hypermarket and so on). Data visualization provides an opportunity for the user to analyze data changes (seasonal volatility, distinctions of products within different stores) in a single click. This analysis is a great tool for decision-making support. Main goals of this particular research are: managerial decisions related to the adjustments of the assortment of different shops, taking in a consideration features of each of the stores AND development of “afigage” for navigation in the supermarkets/hypermarkets, in order to stimulate cross-selling purchases (based on clients’ preferences of different segments).

Created visualization models showed clearly different varieties of customers’ baskets. As for segmentation, it was not the initial purpose of the research, however the success of the division of customers in different groups based on market basket analysis is impressive. More than 88% of

transactions were distributed, this is a sign for effective segmentation. Extra benefit of this approach, as I mentioned before.

Yes, IT IS a tool for profit maximization. Market basket analysis solves three main problems: assortment changes, merchandising and, plus, segmentation. Profit maximization is possible to achieve through adjustment of assortment policies and stimulation of cross-selling purchases.

This approach should be used on a regular basis. It will be useful for identification of different changes within the market baskets. It is also important to react flexibly on all the possible changes and take decisions in good time".

This interview gave comprehensive responses to all the questions. To summarize, market basket visualization helps to solve several issues and be the base for effective decision-making (Figure 22)

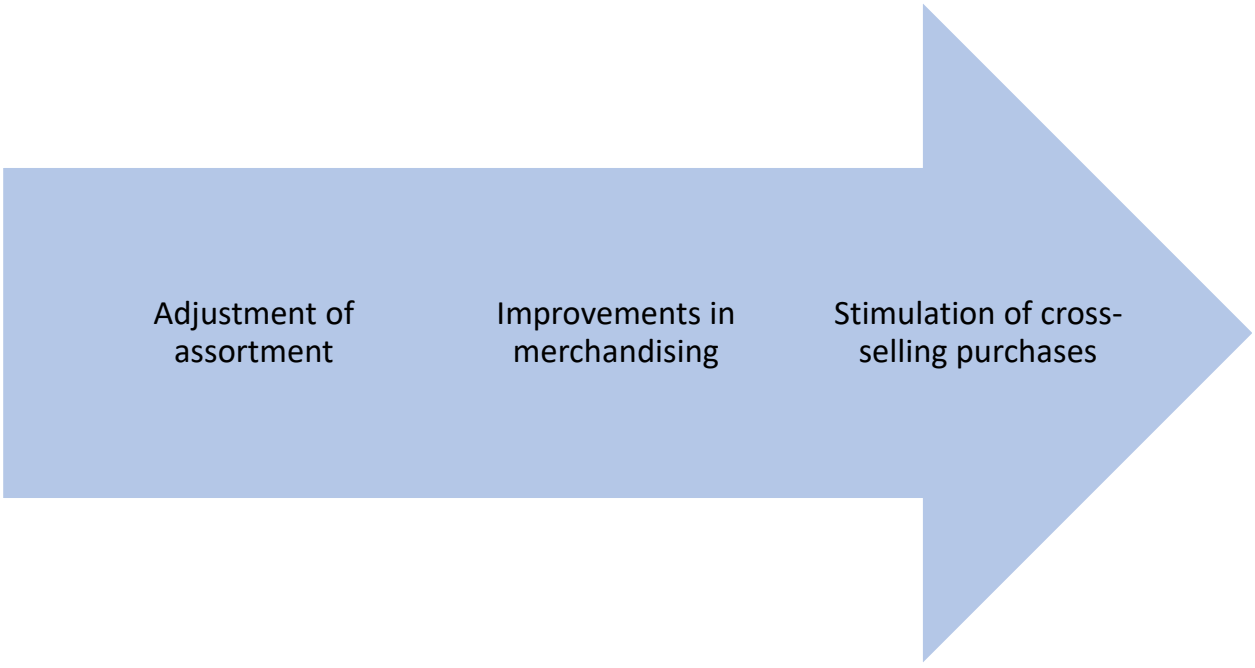

Fig. 22 Implementation of market basket analysis

Market basket analysis is not a newest technique, however, the positive outcome of this analysis is underestimated. Visualization models made this approach work even more effectively. Not only market baskets of different types were successfully obtained, but also extra value was extracted: segmentation of consumers.

3.5 Summary of chapter 3

Market basket visualization with the use of experimental SQL Queries lead to effective outcomes. Before data manipulations in Oracle Data Miner and Oracle Data Visualization were conducted, some transformations had to be made. Initially, there were 683 categories and segments of products. If this amount of categories were united to cross-selling pairs and, later, visualized; visual models would have been extremely cluttered and, as a result, uninformative. That is why some categories and segments were united, based on the analysis of the researcher and in-depth interviews. After transformation of the nomenclatures, 157 segments and categories were obtained.

After these manipulations, experimental SQL Queries were used for market basket analysis. The researcher prepared the assumptions, which were approved by the senior analyst and the manager of the department of the researched company, which had to be explored on "raw" Big Data. Intervals for analysis taken were transactions for two months on two hypermarkets (in total one year). This interval was chosen, in order to increase the speed of data manipulations in Oracle Data Mining and to see the seasonal changes in data during the year. Some tables had millions of rows of transactions. Based on the assumptions, filters in SQL Queries were created, in order to obtain different market baskets. If the share of the particular market basket in the total amount of transactions exceeded 5%, analysis of this customers' basket was analyzed further. Significant market baskets identified in data were also visualized, in order to compare the effectiveness of data mining results and search oh hidden interconnections in data. Visualization models were built based on sample tables from the data that contained exactly 600000 rows of transactions.

Visualization demonstrates clearly the market baskets obtained. Apart from that, it shows important cross-selling pairs, which might not be in a particular customers' basket, however, they stand out in data. Visualization simplifies comprehension of the results; visualized models can be communicated to top managers and based on them important decisions can be made.

Experimental data mining techniques and visualization allowed not only get results about various market baskets, but also segment clients relying on these types of baskets. At the beginning of the research it was not expected to obtain efficient segmentation of customers, however, it became possible. So this analysis provided an extra value to the research.

This approach can be used in practice in the researched company (and not only in that company) on a regular basis. For managers the benefit of this market basket analysis with visualization models is that to support their decisions for adjustment of the assortment and improvement of merchandising. These changes will stimulate cross-selling consumer behavior. The more the clients purchase, the higher is the profit of the organizations. Regularity of this analysis is necessary for flexible reactions on any changes in market baskets.

Conclusion

The goal of this research was to develop a new model that will improve imperfections of the existing explored models, which was met. Main disadvantages of the previous researches devoted to market basket analysis was the absence of visualization, which could help the analyst interpret the data and identify hidden correlations. Another common drawback was related to preliminary data transformation and cluster algorithm as a data mining tool. This type of algorithm was already criticized by different authors several years ago (in 2014 and 2015), however, this method is still used. Cluster analysis is widely-spread for statistical analysis, but it is not applicable for Big Data analysis. That is why experimental approach with the use of SQL Queries was created in this research.

The new model obtained with the efficient data transformation lead to successful and clear visualization. The visualized models compared to the outcomes after data mining process represented the whole picture of the market basket through the Network model. Due to visualization, it was possible not only to identify the whole baskets (not just cross-selling pairs), but also to see the hidden meaningful correlation, that were not visible in tables in Oracle Data Miner before the creation of visualization.

Popular market basket structures were also found with the use of the new model. Overall, 7 different "big" baskets were obtained (which had a significant share in the total amount of transactions, more than 2%). These baskets are: Vegetarian basket, basket for "meat lovers", non-food basket, French basket, Parent's basket, smoker's basket and alcohol basket.

Market basket visualization is an effective tool for decision-making support. It can be implemented by managers also as a tool for adjustment of assortment. Apart from that, this method can be a base for managerial decision about changes in merchandising. Flexible adjustments in assortment and effective positioning of goods in the hypermarkets based on the analysis of customer behavior (which is provided by market basket visualization analysis) together lead to stimulation of cross-selling purchases by consumers, as all these changes fit the preferences of the clients.

As the customers start buying more goods, the sales of the hypermarkets increase, and, as a result, the profit of the organization also goes up. All retail chains face the problem of accrescent volume of data related to transactions, it is needed to analyze this data, in order to extract valuable information about customer behavior. This problem is actual for every retailer. The developed model in this research can be a base for any retailer for exploration of their unique market baskets.

Discussion

The model developed in this research provided another result, apart from identification of the market basket structures and efficient visualization. Another outcome obtained was segmentation of clients based on market basket distribution. 88% of transactions that were analyzed, were correlated to one of the seven baskets that were created during the research. There are intersections in these baskets, however, most of the baskets were made with the use of filters (programmed in SQL Queries), which excluded intersections with other baskets, but still not with all of them. So, overall, the segmentation obtained can be evaluated as effective.

The manager of the Department of Innovations of the researched company also approved the good result of segmentation based on market basket analysis, even though this was unexpected to get. "As for segmentation, it was not the initial purpose of the research, however the success of the division of customers in different groups based on market basket analysis is impressive. More than 88% of transactions were distributed, this is a sign for effective segmentation. Extra benefit of this approach, as I mentioned before" – this is the comment of the manager related to the topic of segmentation.

Further research in this field is required to understand, how effective this approach for segmentation will be on different data sets. It is also necessary to analyze deeper all the possible advantages and disadvantages of this research. Apart from that, comparison analysis with other tools for segmentation should be conducted. There are many more questions that remain unanswered related to the newly obtained method of segmentation based on market basket analysis and segmentation. This issue to be solved in further researches.

References

1. Albionresearch.com. (2018). Data Mining: Market Basket Analysis. [online] Available at: http://www.albionresearch.com/data_mining/market_basket.php [Accessed 21 Mar. 2018].
2. Alonso-Betanzos, A., Gámez, J., Herrera, F., Puerta, J., Riquelme, J. and Laney Douglas (2017). Volume, variety and velocity in Data Science. *Knowledge-Based Systems*, 117, p.16.
3. Berry, M. and Linoff, G. (2011). Data Mining Techniques for Marketing, Sales, and Customer Support. 3rd ed. John Wiley & Sons, pp.3-45.
4. Boztuğ, Y. and Reutterer, T. (2017). A combined approach for segment-specific market basket analysis. European Journal of Operational Research, 187(1), pp.294-312.
5. Bradlow, E., Gangwar, M., Kopalle, P. and Voleti, S. (2017). The Role of Big Data and Predictive Analytics in Retailing. Journal of Retailing, 93(1), pp.79-95.
6. Bramer, M. (2013). Principles of Data Mining. 2nd ed. Springer, p.5.
7. Brath, R. and Jonker, D. (2015). Graph analysis and visualization. Indianapolis: John Wiley and Sons, pp.3-9.
8. BusinessDictionary.com. (2018). What comes after those ellipses?. [online] Available at: http://www.businessdictionary.com/definition/hypothesis.html [Accessed 22 Mar. 2018].
9. Cios, K., Pedrycz, W., Swiniarski, R. and Kurgan, L. (2007). Data Mining. New York: Springer, pp.9-23.
10. Constine, J. (2017). *How Big Is Facebook's Data? 2.5 Billion Pieces Of Content And 500+ Terabytes Ingested Every Day*. [online] TechCrunch. Available at: https://techcrunch.com/2012/08/22/how-big-is-facebooks-data-2-5-billion-pieces-of-content-and-500-terabytes-ingested-every-day/ [Accessed 7 May 2017].
11. Cooper, D. and Schindler, P. (2006). Business Research Methods. 9th ed. New York: McGraw Hill International Edition, pp. 144, 185, 196, 198-200.

12. Coyle, K. (2006). Mass Digitization of Books. *The Journal of Academic Librarianship*, 32(6), pp.641-645.
13. Craven, M. and Page, C. (2015). Big Data in Healthcare: Opportunities and Challenges. *Big Data*, 3(4), pp.209-210.
14. De Mauro, A., Greco, M. and Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. *Library Review*, 65(3), pp.122-135.
15. Diffen.com. (2017). *Data vs Information - Difference and Comparison | Diffen*. [online] Available at: http://www.diffen.com/difference/Data_vs_Information [Accessed 7 May 2017].
16. En.wikipedia.org. (2017). *Data visualization*. [online] Available at: https://en.wikipedia.org/wiki/Data_visualization#cite_note-MF08-1 [Accessed 7 May 2018].
17. Friendly, M. and Denis, D. (n.d.). *Milestones in the history of thematic cartography, statistical graphics, and data visualization*. 1st ed. [Toronto, Ont.]: [Statistical Consulting Service, York University], pp.118-121.
18. Ghauri, P. and Grønhaug, K. (2005). Research methods in business studies. 3rd ed. Harlow, England: Financial Times Prentice Hall, p.100.
19. Gutierrez, D. (2018). How Retailers are Using Big Data to Their Advantage - insideBIGDATA. [online] insideBIGDATA. Available at: https://insidebigdata.com/2017/01/21/how-retailers-are-using-big-data-to-their-advantage/ [Accessed 21 Mar. 2018].
20. Hassani, M. and Seidl, T. (2016). Clustering Big Data streams: recent challenges and contributions. it - Information Technology, 58(4), pp.39-54.
21. Hilbert, M. (2015). Big Data for Development: A Review of Promises and Challenges. *Development Policy Review*, 34(1), pp.135-174.
22. Insights, S., Insights, B. and Data?, W. (2017). *What is Big Data and why it matters*. [online] Sas.com. Available at: https://www.sas.com/en_us/insights/big-data/what-is-big-data.html [Accessed 7 May 2018].

23. Kim, T. and Wang, H. (2013). Special issue on Assembly Technologies and Systems – selected papers from the 4th CIRP Conference on Assembly Technologies and Systems. *Journal of Manufacturing Systems*, 32(3), p.403.
24. Kurasova, O., Medvedev, V. and Stefanovic, P. (2014). Strategies for Big Data Clustering. 2014 IEEE 26th International Conference on Tools with Artificial Intelligence, 740-747, p.745.
25. Manco, G., Rullo, P., Gallucci, L. and Paturzo, M. (2016). Rialto: A Knowledge Discovery suite for data analysis. Expert Systems with Applications, 59, pp.145-164.
26. Mark, M. (2011). *Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data*. 28th ed.
27. Mayer-Schönberger, V. and Cukier, K. (2013). *Big data*. 1st ed. Boston, Mass.: Houghton Mifflin, pp.2-4, 6-8.
28. McKinsey & Company. (2017). *Big data: The next frontier for innovation, competition, and productivity*. [online] Available at: http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation [Accessed 7 May 2018].
29. Mirkes, E., Coats, T., Levesley, J. and Gorban, A. (2016). Handling missing data in large healthcare dataset: A case study of unknown trauma outcomes. *Computers in Biology and Medicine*, 75, pp.203-216.
30. Mohd Shaid, S. and Maarof, M. (2014). Malware Behaviour Visualization. *Jurnal Teknologi*, 70(5).
31. Saunders, M., Lewis, P. and Thornhill, A. (2016). Research methods for business students. 7th ed. Pearson Education Limited, pp.171, 389.
32. SearchBusinessAnalytics. (2018). What is association rules (in data mining)? - Definition from WhatIs.com. [online] Available at: http://searchbusinessanalytics.techtarget.com/definition/association-rules-in-data-mining [Accessed 24 Mar. 2018].
33. Shneiderman, B. (2014). The Big Picture for Big Data: Visualization. *Science*, 343(6172), pp.730-730.

34. SURVEY PAPER ON STRATEGIES FOR BIG DATA ANALYSIS AND CLUSTERING. (2017). International Journal of Advance Engineering and Research Development, 4(10).
35. Taft, M., Krishnan, K., Hornick, M., Muhkin, D., Tang, G., Thomas, S. and Stengard, P. (2005). [online] Docs.oracle.com. Available at: https://docs.oracle.com/cd/B19306_01/datamine.102/b14339.pdf [Accessed 19 Apr. 2018].
36. tutor2u. (2017). *ICT: The difference between data and information | tutor2u Business*. [online] Available at: https://www.tutor2u.net/business/reference/the-difference-between-data-and-information [Accessed 7 May 2018].
37. Villanovau.com. (2017). *Cite a Website - Cite This For Me*. [online] Available at: https://www.villanovau.com/resources/bi/what-is-big-data/#.WQ8cYuXyjIU [Accessed 7 May 2018].
38. Zanin, M., Papo, D., Sousa, P., Menasalvas, E., Nicchi, A., Kubik, E. and Boccaletti, S. (2016). Combining complex networks and data mining: Why and how. Physics Reports, 635, pp.1-44.
39. Zerhari, B., Lahcen, A. and Mouline, S. (2015). Big Data Clustering: Algorithms and Challenges. [online] ResearchGate. Available at: https://www.researchgate.net/publication/276934256_Big_Data_Clustering_Algorithms_and_Challenges [Accessed 27 Apr. 2018].
40. Zhang, J. and Huang, M. (2016). Data behaviours model for Big Data visual analytics. *International Journal of Big Data Intelligence*, 3(1), p.2.

Appendix 1. Fragment of program code in SQL Queries

```
select TRAN_ID,CATEGORY_CODE, FAMDSC,
case
when  category_code=147 or category_code=150 then 'Белые хлебы об.'
when category_code=760 or category_code=769 or category_code=779 or category_code=790 or category_code=785 or category_code=781 or category_code=770 or category_code=766 or category_code=782 or category_code=784 then 'об.Прочие бытовые принадл. и средства'
when category_code=250 or category_code=247 or category_code=232 or category_code=246 or category_code=242 or category_code=231 or category_code=238 or category_code=244 or category_code=241 then 'об.Сыры производственные'
when category_code=205 or category_code=197 or category_code=171 or category_code=256 or category_code=173 or category_code=227 or category_code=206 or category_code=229 or category_code=179 or category_code=172 then 'об.Французская выпеч.+заморож.фр.выпеч.'
when category_code=234 or category_code=176 or category_code=237 or category_code=181 or category_code=236 or category_code=178 then 'об. Колбасы / ветчины'
when category_code=481 or category_code=482 or category_code=484 or category_code=477 or category_code=499 or category_code=496 or category_code=485 then 'об. Конфеты и леденцы'
when category_code=151 or category_code=132 or category_code=133 or category_code=122 or category_code=131 or category_code=157 or category_code=124 or category_code=158 then 'об. Парфюмерия и косметика'
when category_code=762 or category_code=771 or category_code=772 or category_code=759 or category_code=765 or category_code=768 then 'об. Школьные и канцелярские товары'
when category_code=668 or category_code=680 or category_code=677 or category_code=679 then 'об. Вода'
when category_code=681 or category_code=690 or category_code=977 or category_code=974 or category_code=160 or category_code=691 or category_code=693 or category_code=678 or category_code=684 or category_code=961 then 'об.Садоводство'
when category_code=492 or category_code=682  or category_code= 495  or category_code=683 or category_code= 504 or category_code= 501 or category_code= 498  or category_code=507  or category_code=685  or category_code=688 then 'об. Корма /товары для дом.животных'
when category_code=583 or category_code=587 or category_code=586 or category_code=161 or category_code=585 or category_code=162 or category_code=163 or category_code=665 or category_code=168 or category_code=164 or category_code=589 or category_code=588 or category_code=202 then 'об. Злак. и диетич. батончики'
```

when category_code=146 or category_code=139 or category_code=145 or category_code=149 or category_code=137 or category_code=143 or category_code=155 or category_code=140 or category_code=142 or category_code=117 or category_code=129 or category_code=111 or category_code=123 or category_code=120 or category_code=126 or category_code=114 or category_code=121 or category_code=136 then 'об. Мясо рублен.'
when category_code=748 or category_code=747 or category_code=744 or category_code=749 then 'об. Средства для стирки'
when category_code=291 or category_code=289 or category_code=290 or category_code=292 then 'об. Гот-е к употр/охлаждён. мореп.'
when category_code=814 or category_code=834 or category_code=816 or category_code=815 or category_code=817 or category_code=837 or category_code=838 or category_code=839 or category_code=840 then 'об. Газеты и журналы'
when category_code=267 or category_code=190 or category_code=233 or category_code=264 or category_code=235 or category_code=177 or category_code=239 or category_code=186 or category_code=240 or category_code=192 or category_code=265 then 'об. Деликатесы и мясопродукты'
when category_code=279 or category_code=282 or category_code=278 then 'об. Сыры и молочные продукты (магазин. упак)'
when category_code=510 or category_code=516 or category_code=519 or category_code=513 or category_code=522 or category_code=525 then 'об. Авто/мото принадл.'
when category_code=260 or category_code=258 or category_code=262 or category_code=261 or category_code=253 or category_code=259 or category_code=275 or category_code=269 or category_code=266 then 'об. Заморож. блюда, акссесуар.'
when category_code=647 or category_code=633 or category_code=651 then 'об. Кухонные аксесс.'
when category_code=742 or category_code=567 or category_code=564 or category_code=729 or category_code=738 or category_code=728 or category_code=739 or category_code=735 then 'об. Гигиена для малышей'
when category_code=801 or category_code=786 or category_code=848 or category_code=788 or category_code=795 or category_code=846 or category_code=787 or category_code=789 or category_code=780 or category_code=849 or category_code=798 or category_code=792 or category_code=777 then 'об. Игрушки'
when category_code=187 or category_code=112 or category_code=130 or category_code=141 then 'об. Птица прочая'
when category_code=303 or category_code=306 then 'об. Фрукты прочие'

when category_code=480 or category_code=643 or category_code=483 or category_code=641 or category_code=655 or category_code=644 or category_code=486 or category_code=476 then 'об. Вино'
when category_code=694 or category_code=727 or category_code=719 or category_code=695 or category_code=718 or category_code=689 then 'об. Питание для детей'
when category_code=417 or category_code=402 or category_code=434 then 'об. Еда быстрого пригот.'
when category_code=661 or category_code=660 or category_code=662 then 'об. Хозяйственные аксс.'
when category_code=302 or category_code=301 or category_code=304 or category_code=305 then 'об. Кейтеринг и охл.закуски'
when category_code=274 or category_code=272 then 'об. Инд. мороженые батончики (=мороженое)'
when category_code=913 or category_code=911 or category_code=910 then 'об. Муж. нижнее бельё и носки'
when category_code=883 or category_code=885 then 'об. Колготки, чулки, носки жен.'
when category_code=669 or category_code=666 or category_code=674 or category_code=671 or category_code=989 or category_code=676 or category_code=979 then 'об. Товары для ремонта'
when category_code=810 or category_code=813 or category_code=807 then 'об. Книги'
when category_code=636 or category_code=663 or category_code=664 then 'об. Сервировочная посуда'
when category_code=469 or category_code=470 or category_code=478 then 'об. Водка натуральная'
when category_code=127 or category_code=125 or category_code=128 then 'об. прочие средства для волос'
when category_code=868 or category_code=931 or category_code=854 or category_code=856 or category_code=857 or category_code=855 or category_code=869 or category_code=864 or category_code=858 or category_code=867 or category_code=865 then 'об. обувь жен/муж/дет'
when category_code=642 or category_code=640 or category_code=645 or category_code=646 then 'об. текстиль и декор'
when category_code=590 or category_code=598 or category_code=632 then 'об. алкогольные напитки и аперитивы'
when category_code=993 or category_code=987 or category_code=990 or category_code=988 or category_code=991 or category_code=994 or category_code=997 then 'об. нижнее женское белье'

when category_code=634 or category_code=946 or category_code=947 or category_code=945 then 'об. посуда для кухни'
when category_code=986 or category_code=908 or category_code=958 or category_code=985 or category_code=959 or category_code=984 or category_code=942 or category_code=962 or category_code=965 or category_code=917 then 'об. детское ниж белье,колготки,носки'
when category_code=654 or category_code=657 then 'об. лампы и приборы для освещения'
when category_code=457 or category_code=463 or category_code=464 or category_code=466 or category_code=460 then 'об. соленые закуски'
when category_code=194 or category_code=148 then 'об. черный хлеб'
when category_code=710 or category_code=686 or category_code=978 or category_code=746 or category_code=843 or category_code=753 or category_code=716 or category_code=713 or category_code=715 or category_code=743 then 'об. спорт товары'
when category_code=382 or category_code=337 or category_code=373 or category_code=332 or category_code=319 or category_code=338 or category_code=331 or category_code=329 or category_code=328 or category_code=326 or category_code=322 or category_code=313 or category_code=312 or category_code=311 or category_code=310 or category_code=308 or category_code=307 then 'об. бакалея на развес'
when category_code=907 or category_code=905 or category_code=904 or category_code=914 then 'об. галантерея и аксессуары'
when category_code=273 or category_code=277 or category_code=270 then 'об. растения и цветы'
when category_code=285 or category_code=294 then 'об. рыба охлажденная и рыбные продукты'
when category_code=316 or category_code=330 or category_code=333 or category_code=325 or category_code=324 or category_code=327 or category_code=336 or category_code=334 then 'об. гриль, готовая продукция'
when category_code=106 or category_code=897 or category_code=900 or category_code=943 then 'об. гигиена прочие товары'
when category_code=299 or category_code=134 then 'об. рыба копченая'
when category_code=109 or category_code=107 then 'об. макияж, одеколоны и духи'
when category_code=416 or category_code=409 or category_code=413 or category_code=397 or category_code=389 or category_code=375 or category_code=387 or category_code=361 or category_code=360 or category_code=362 or category_code=442 or category_code=367 or category_code=355 or category_code=363 or category_code=407 or category_code=370 or category_code=356 or category_code=411 or category_code=341 or category_code=368 or

category_code=369 or category_code=412 or category_code=340 or category_code=410 or category_code=371 or category_code=406 or category_code=365 or category_code=366 or category_code=344 or category_code=388 or category_code=359 or category_code=390 then 'об. электробытовые приборы малых и средних размеров'
when category_code=628 or category_code=631 or category_code=629 or category_code=627 or category_code=630 then 'об. телефония и аксессуары д/тел'
when category_code=705 or category_code=707 or category_code=711 or category_code=706 or category_code=703 or category_code=708 or category_code=709 or category_code=704 or category_code=717 or category_code=714 or category_code=702 then 'об. мужская одежда'
when category_code=638 or category_code=637 then 'об. для стола аксессуары'
when category_code=950 or category_code=953 or category_code=951 or category_code=952 or category_code=700 then 'об. товары для отдыха на открытом воздухе'
when category_code=756 or category_code=755 or category_code=764 or category_code=773 or category_code=758 or category_code=774 or category_code=757 or category_code=783 then 'об. одежда для новорожденных и детей до 3-х лет'
when category_code=723 or category_code=736 or category_code=722 or category_code=732 or category_code=737 or category_code=741 or category_code=752 or category_code=797 or category_code=726 or category_code=751 or category_code=750 or category_code=721 or category_code=731 or category_code=730 or category_code=720 then 'об. одежда для детей'
when category_code=871 or category_code=870 or category_code=876 or category_code=879 then 'об. обувь для дома муж/жен/дет'
when category_code=196 or category_code=154 or category_code=188 or category_code=152 then 'об. хлебные производные'
when category_code=616 or category_code=619 or category_code=626 or category_code=582 or category_code=580 or category_code=581 or category_code=620 or category_code=604 or category_code=622 or category_code=617 or category_code=610 or category_code=614 or category_code=623 or category_code=576 or category_code=575 or category_code=605 or category_code=593 or category_code=577 or category_code=613 or category_code=579 or category_code=607 or category_code=578 or category_code=602 or category_code=611 or category_code=625 or category_code=608 or category_code=584 then 'об. аудио/видео техника и аксесс.'
when category_code=533 or category_code=531 or category_code=529 or category_code=511 or category_code=512 or category_code=527 or category_code=530 or category_code=515 or category_code=536 or category_code=520 or category_code=521 or category_code=514 or category_code=532 or category_code=503 or category_code=523 or category_code=524 or

category_code=506 or category_code=535 or category_code=517 or category_code=502 or category_code=528 or category_code=509 or category_code=505 or category_code=534 or category_code=526 or category_code=518 or category_code=500 then 'об. пк и комплектующие к пк'
when category_code=673 or category_code=672 or category_code=934 or category_code=692 or category_code=600 or category_code=699 or category_code=601 or category_code=701 or category_code=687 or category_code=796 or category_code=603 then 'об. Женска одежда'
when category_code=697 or category_code=998 then 'об. Украшение для праздника'
when category_code=549 or category_code=659 or category_code=546 or category_code=658 then 'об. Принадлеж. для новорожденных'
when category_code=835 or category_code=832 or category_code=833 or category_code=804 or category_code=831 or category_code=803 or category_code=805 or category_code=836 or category_code=851 or category_code=852 or category_code=853 or category_code=802 then 'об. Видео- аудио- носители'
when category_code=648 or category_code=653 or category_code=650 or category_code=652 or category_code=649 then 'об. Покрытия для пола, стен, потолка'
when category_code=800 or category_code=591 or category_code=818 or category_code=824 or category_code=594 or category_code=821 or category_code=830 or category_code=820 or category_code=827 or category_code=799 then 'об. Мебель'
when category_code=775 or category_code=776 or category_code=845 or category_code=860 then 'об. Сувениры, подарки'
when category_code=906 or category_code=912 or category_code=903 or category_code=901 or category_code=938 or category_code=924 or category_code=909 or category_code=902 or category_code=937 or category_code=936 or category_code=935 or category_code=933 or category_code=932 or category_code=930 or category_code=929 or category_code=928 or category_code=927 or category_code=926 or category_code=916 or category_code=915 then 'об. Органические продукты'
when category_code=559 or category_code=540 or category_code=553 or category_code=551 or category_code=544 or category_code=556 or category_code=561 or category_code=550 or category_code=562 or category_code=542 or category_code=541 or category_code=545 then 'об. Порт. техника'
when category_code=448 or category_code=424 or category_code=449 or category_code=445 or category_code=423 or category_code=428 or category_code=443 or category_code=422 or category_code=427 or category_code=425 or category_code=421 or category_code=430 or category_code=446 then 'об. Электробытовая крупная техника'

```
when category_code=563 or category_code=565 or category_code=548 or category_code=569 or category_code=574 or category_code=557 or category_code=547 or category_code=566 or category_code=572 or category_code=571 or category_code=568 or category_code=560 then 'об.Аксес. и услуги для фото '
when category_code=169 or category_code=165 or category_code=175 or category_code=170 or category_code=167 or category_code=166 then 'об. Пирожные, торты сырьё'
when category_code=822 or category_code=825 or category_code=819 or category_code=828 then 'об. Игры и игровые приставки'
when category_code=596 or category_code=595 or category_code=592 then 'об. Кухни и акссес.'
when category_code=981 or category_code=980 or category_code=976 or category_code=972 or category_code=970 or category_code=967 or category_code=966 or category_code=964 or category_code=963 or category_code=954 or category_code=949 then 'об. Столовые приборы, ювелирка, часы'


else FAMDSC
END as AGGREGAT_CATEGORY
from "SQL join_N$10032"
```

Appendix 2. Fragment of the output after use of built-in function "Association"

ID	Antecedent	Consequent	Lift	Confidence(%)	Support(%)	Item Count
999	300	481	4.821	22.465	2.526	1
1602	558	554	5.427	38.379	3.423	1
1580	558	552	6.205	32.661	2.913	1
5297	791	779	4.77	12.643	1.126	1
5301	791	790	5.663	13.139	1.17	1
5299	791	785	6.457	13.122	1.169	1
5353	939	941	5.1	39.381	3.376	1
5085	570	573	5.561	42.748	3.574	1
1460	635	455	5.685	15.696	1.28	1
1311	454	402	6.543	19.582	1.527	1
5354	941	939	5.1	43.716	3.376	1
56	941	135	4.826	26.398	2.038	1
5086	573	570	5.561	46.504	3.574	1
1369	447	552	5.024	26.446	1.958	1
1373	447	555	4.798	16.86	1.248	1
700	108	881	5.507	35.012	2.549	1
710	108	944	5.338	16.16	1.177	1
1601	554	558	5.427	48.409	3.423	1
1578	554	552	7.346	38.667	2.734	1
2917	288	287	5.376	19.478	1.333	1
7191	193 AND 201	287	4.739	17.171	1.165	2
5216	690	681	5.713	32.712	2.176	1
9345	309 AND 968	969	5.027	57.627	3.831	2
1321	420	508	4.765	20.61	1.348	1
1329	420	555	5.289	18.587	1.216	1
701	881	108	5.507	40.093	2.549	1
9346	309 AND 969	968	4.878	61.328	3.831	2
12580	309 AND 971	318	4.725	68.165	4.194	2
3587	138 AND 193	117	4.798	17.65	1.061	2

1059	302	315	4.785	17.589	1.042	1
7908	193 AND 318	971	4.778	54.602	3.177	2
5215	681	690	5.713	38.005	2.176	1
8178	193 AND 968	969	5.25	60.19	3.308	2
55	135	941	4.826	37.271	2.038	1
5274	767	793	6.547	31.765	1.717	1
5231	767	748	4.779	22.86	1.235	1
5253	767	761	7.043	27.647	1.494	1
5270	767	778	6.248	22.475	1.215	1
5249	767	760	7.704	23.623	1.277	1
5268	767	769	7.537	20.956	1.133	1
9526	318 AND 969	968	4.737	59.548	3.203	2
9534	318 AND 969	971	5.103	58.323	3.137	2
2874	286	287	6.845	24.803	1.326	1
9525	318 AND 968	969	5.235	60.018	3.203	2
9528	318 AND 968	971	5.141	58.759	3.136	2
1579	552	558	6.205	55.347	2.913	1
1370	552	447	5.024	37.2	1.958	1
1577	552	554	7.346	51.949	2.734	1
1499	492	682	10.61	22.742	1.184	1
9018	198 AND 968	969	5.029	57.655	2.991	2
8200	203 AND 198	200	4.719	51.205	2.623	2
8179	193 AND 969	968	5.165	64.925	3.308	2
5265	762	771	9.104	30.584	1.55	1
4356	138 AND 318	971	4.823	55.123	2.708	2
6402	201 AND 318	971	4.851	55.436	2.708	2
3576	110 AND 968	969	4.96	56.864	2.775	2
6585	201 AND 968	969	5.356	61.407	2.989	2
5275	793	767	6.547	35.382	1.717	1
5235	793	748	5.58	26.694	1.295	1
5257	793	761	5.363	21.053	1.021	1
5293	793	778	5.966	21.46	1.041	1
5230	748	767	4.779	25.827	1.235	1
5234	748	793	5.58	27.074	1.295	1

5225	748	744	8.703	21.577	1.032	1
9019	198 AND 969	968	4.99	62.736	2.991	2
9802	968 AND 971	969	5.257	60.27	2.858	2
1000	481	300	4.821	54.21	2.526	1
7945	193 AND 539	420	4.951	32.386	1.507	2
8020	193 AND 539	508	5.322	23.021	1.071	2
9109	200 AND 209	203	4.888	50.226	2.332	2
9536	969 AND 971	318	4.714	68.007	3.137	2
9803	969 AND 971	968	4.929	61.959	2.858	2
4485	138 AND 968	969	5.23	59.961	2.75	2
6586	201 AND 969	968	5.314	66.806	2.989	2
8005	193 AND 558	454	4.758	37.092	1.633	2
7978	193 AND 558	447	5.396	39.957	1.76	2
8074	193 AND 558	554	6.262	44.287	1.95	2
7705	193 AND 558	288	4.719	32.29	1.422	2
7951	193 AND 558	420	5.043	32.992	1.453	2
8065	193 AND 558	552	7.421	39.065	1.72	2
3577	110 AND 969	968	5.035	63.299	2.775	2
1322	508	420	4.765	31.17	1.348	1
4486	138 AND 969	968	5.148	64.712	2.75	2
12018	209 AND 968	969	4.906	56.244	2.378	2
7060	193 AND 203	200	5.002	54.275	2.293	2
12429	300 AND 309	481	5.63	26.237	1.096	2
6945	209 AND 300	481	6.229	29.026	1.207	2
5940	200 AND 318	971	4.939	56.442	2.345	2
12138	210 AND 318	971	4.821	55.097	2.221	2
11330	193 AND 309 AND 198	200	4.741	51.445	2.031	3
5252	761	767	7.043	38.062	1.494	1
5256	761	793	5.363	26.021	1.021	1
5247	761	760	9.366	28.721	1.127	1
9438	318 AND 539	971	5.052	57.734	2.263	2
7981	193 AND 791	447	4.752	35.182	1.342	2
8143	193 AND 791	767	5.067	27.386	1.045	2

8145	193 AND 791	793	5.67	27.511	1.049	2
9244	309 AND 558	554	5.837	41.279	1.574	2
9235	309 AND 558	552	6.571	34.587	1.319	2
6759	203 AND 318	971	4.884	55.819	2.125	2
7977	193 AND 447	558	5.251	46.84	1.76	2
7965	193 AND 447	454	5.441	42.421	1.594	2
7974	193 AND 447	554	5.089	35.99	1.352	2
7693	193 AND 447	288	5.127	35.08	1.318	2
7939	193 AND 447	420	4.973	32.531	1.222	2
7971	193 AND 447	552	6.225	32.767	1.231	2
9276	309 AND 570	573	6.076	46.699	1.746	2
7740	193 AND 300	481	6.014	28.025	1.045	2
13801	198 AND 318 AND 309	971	5.053	57.748	2.141	3
5979	200 AND 968	969	5.208	59.711	2.211	2
6439	201 AND 539	447	4.725	34.986	1.291	2
6424	201 AND 539	420	4.74	31.01	1.144	2
14513	209 AND 318 AND 309	971	4.965	56.743	2.087	3
1060	315	302	4.785	28.347	1.042	1
11310	193 AND 201 AND 198	200	4.756	51.604	1.892	3
13587	198 AND 309 AND 209	203	4.801	49.339	1.807	3
7626	193 AND 284	288	5.163	35.324	1.292	2
9038	201 AND 203	200	4.893	53.095	1.935	2
8887	198 AND 558	447	4.87	36.06	1.313	2
8953	198 AND 558	554	6.055	42.819	1.559	2
8947	198 AND 558	552	6.903	36.337	1.323	2
11976	209 AND 570	573	6.546	50.315	1.823	2
2916	287	288	5.376	36.782	1.333	1
2875	287	286	6.845	36.583	1.326	1
13173	193 AND 318 AND 309	971	5.136	58.699	2.123	3

6448	201 AND 558	447	5.584	41.345	1.494	2
6520	201 AND 558	554	6.282	44.427	1.605	2
6259	201 AND 558	288	4.82	32.976	1.191	2
6427	201 AND 558	420	4.825	31.564	1.14	2
6511	201 AND 558	552	7.527	39.619	1.432	2
8004	193 AND 454	558	5.071	45.236	1.633	2
7966	193 AND 454	447	5.96	44.133	1.594	2
8001	193 AND 454	554	5.024	35.526	1.283	2
7942	193 AND 454	420	5.396	35.3	1.275	2
7998	193 AND 454	552	5.887	30.986	1.119	2
5271	778	767	6.248	33.767	1.215	1
5292	778	793	5.966	28.945	1.041	1
14689	309 AND 968 AND 318	969	5.507	63.131	2.241	3
14693	309 AND 968 AND 318	971	5.449	62.277	2.211	3
12282	279 AND 318	971	4.944	56.504	2.001	2
14690	309 AND 969 AND 318	968	5.042	63.386	2.241	3
14697	309 AND 969 AND 318	971	5.429	62.045	2.193	3
1374	555	447	4.798	35.525	1.248	1
1330	555	420	5.289	34.602	1.216	1
11950	209 AND 558	554	5.781	40.883	1.427	2
11947	209 AND 558	552	6.33	33.32	1.163	2
5980	200 AND 969	968	5.084	63.918	2.211	2
4387	138 AND 558	447	5.1	37.76	1.303	2
4441	138 AND 558	554	6.1	43.14	1.489	2
4435	138 AND 558	552	7.138	37.574	1.297	2
12183	210 AND 968	969	5.111	58.595	2.01	2
9641	554 AND 558	454	5.02	39.135	1.34	2
9602	554 AND 558	447	5.495	40.688	1.393	2
9560	554 AND 558	420	4.757	31.119	1.065	2
9695	554 AND 558	552	9.756	51.353	1.758	2

8115	193 AND 570	573	6.451	49.588	1.677	2
3599	939 AND 941	135	5.992	32.773	1.106	2
11977	209 AND 573	570	6.469	54.095	1.823	2
13299	198 AND 309 AND 201	200	4.825	52.349	1.76	3
5266	771	762	9.104	46.149	1.55	1
7950	193 AND 420	558	4.849	43.258	1.453	2
7941	193 AND 420	454	4.868	37.951	1.275	2
7938	193 AND 420	447	4.914	36.386	1.222	2
7704	193 AND 288	558	4.761	42.467	1.422	2
7692	193 AND 288	447	5.315	39.356	1.318	2
13330	198 AND 309 AND 200	203	4.807	49.399	1.653	3
8073	193 AND 554	558	6.548	58.41	1.95	2
8002	193 AND 554	454	4.928	38.42	1.283	2
7975	193 AND 554	447	5.469	40.492	1.352	2
7702	193 AND 554	288	4.857	33.233	1.11	2
7948	193 AND 554	420	4.869	31.851	1.063	2
8062	193 AND 554	552	8.658	45.574	1.522	2
13255	193 AND 969 AND 968	447	4.98	36.875	1.22	3
9684	539 AND 968	969	5.207	59.694	1.971	2
8979	198 AND 570	573	6.147	47.249	1.559	2
13829	198 AND 968 AND 309	969	5.36	61.445	2.02	3
6451	201 AND 791	447	4.788	35.45	1.16	2
9277	309 AND 573	570	6.435	53.806	1.746	2
12627	309 AND 447	454	4.859	37.885	1.229	2
6447	201 AND 447	558	5.168	46.096	1.494	2
6435	201 AND 447	454	5.288	41.223	1.336	2
6444	201 AND 447	554	4.982	35.235	1.142	2
6253	201 AND 447	288	5.114	34.987	1.134	2
6421	201 AND 447	420	4.737	30.987	1.004	2
6441	201 AND 447	552	6.28	33.055	1.071	2

13209	193 AND 968 AND 309	969	5.54	63.513	2.054	3
12184	210 AND 969	968	4.955	62.291	2.01	2
9324	309 AND 939	941	5.204	40.186	1.296	2
8163	193 AND 939	941	5.708	44.079	1.415	2
14285	201 AND 318 AND 309	971	5.189	59.3	1.9	3
14761	318 AND 969 AND 968	971	5.439	62.157	1.991	3
12628	309 AND 454	447	5.208	38.562	1.229	2
10441	110 AND 318 AND 309	971	5.026	57.442	1.805	3
14763	318 AND 971 AND 969	968	5.048	63.462	1.991	3
14762	318 AND 971 AND 968	969	5.538	63.488	1.991	3
10593	138 AND 201 AND 193	447	4.92	36.429	1.142	3
11249	138 AND 318 AND 309	971	5.167	59.05	1.844	3
9243	309 AND 554	558	5.653	50.427	1.574	2
9232	309 AND 554	552	7.532	39.649	1.237	2
9685	539 AND 969	968	5.073	63.775	1.971	2
12012	209 AND 939	941	5.398	41.686	1.286	2
9747	558 AND 968	969	5.614	64.364	1.984	2
9468	318 AND 558	971	4.773	54.545	1.682	2
9644	558 AND 968	454	4.758	37.096	1.144	2
9608	558 AND 968	447	5.535	40.985	1.264	2
9719	558 AND 968	554	6.089	43.062	1.328	2
9448	318 AND 558	554	5.807	41.064	1.266	2
9707	558 AND 968	552	7.532	39.649	1.222	2
9445	318 AND 558	552	6.378	33.574	1.035	2
13830	198 AND 969 AND 309	968	5.217	65.589	2.02	3

11629	193 AND 318 AND 198	971	5.051	57.721	1.775	3
11334	193 AND 318 AND 198	200	4.775	51.813	1.594	3
9543	335 AND 968	969	4.906	56.247	1.727	2
14695	309 AND 971 AND 968	318	4.997	72.085	2.211	3
14742	309 AND 971 AND 968	969	5.578	63.945	1.961	3
5248	760	767	7.704	41.632	1.277	1
5246	760	761	9.366	36.768	1.127	1
8886	198 AND 447	558	4.803	42.844	1.313	2
8877	198 AND 447	454	5.091	39.692	1.217	2
8883	198 AND 447	554	4.784	33.833	1.037	2
13210	193 AND 969 AND 309	968	5.342	67.157	2.054	3
13802	198 AND 971 AND 309	318	4.898	70.65	2.141	3
13411	198 AND 971 AND 309	200	4.734	51.369	1.557	3
711	944	108	5.338	38.863	1.177	1
8116	193 AND 573	570	6.643	55.544	1.677	2
9623	539 AND 558	454	5.05	39.369	1.18	2
9587	539 AND 558	447	5.395	39.943	1.197	2
9667	539 AND 558	554	6.607	46.725	1.401	2
9557	539 AND 558	420	5.606	36.675	1.099	2
9664	539 AND 558	552	7.332	38.594	1.157	2
14309	201 AND 968 AND 309	969	5.621	64.439	1.931	3
1310	402	454	6.543	51.015	1.527	1
14339	201 AND 969 AND 968	447	5.055	37.427	1.119	3
6771	203 AND 968	969	4.978	57.075	1.702	2
9140	203 AND 968	200	4.819	52.286	1.56	2

14699	309 AND 971 AND 969	318	5.103	73.606	2.193	3
14743	309 AND 971 AND 969	968	5.234	65.803	1.961	3
8878	198 AND 454	447	5.528	40.935	1.217	2
8904	198 AND 454	554	4.823	34.107	1.014	2
11665	193 AND 968 AND 198	969	5.466	62.668	1.86	3
5568	100 AND 968	969	4.81	55.148	1.636	2
3547	110 AND 558	554	5.631	39.821	1.179	2
3544	110 AND 558	552	6.431	33.853	1.002	2
12701	193 AND 209 AND 201	210	4.744	52.483	1.551	3
13174	193 AND 971 AND 309	318	4.987	71.933	2.123	3
13218	193 AND 971 AND 309	969	4.77	54.688	1.614	3
8952	198 AND 554	558	5.932	52.913	1.559	2
8884	198 AND 554	447	4.752	35.189	1.037	2
8944	198 AND 554	552	7.902	41.596	1.226	2
13331	198 AND 309 AND 203	200	5.174	56.142	1.653	3

Appendix 3. Example of questions for in-depth interviews

Question	Answer
Is it possible to use built-in functions in Oracle Data Miner for effective Big Data analysis?	No. Even though good tools for Big Data analysis, such as Oracle IT products, have powerful functions, analysts should actively use SQL Queries, re-check the outputs. Working with Big Data is safer when you use programing.
What mistakes can be made during the pre-processing of data in this particular case (market basket analysis)?	Forgetting to filter out trash categories at the very beginning of the data transformation process can lead to huge amount mistakes in the output. And it might not be obvious that the result is wrong. This is the main problem.
Will it be meaningful to analyze transactions that contain meat products (create filter that selects all transactions with meat categories) without any additional filters?	Yes, it is a good idea to have a look at preliminary results. If the share of the transactions in total amount will be meaningful, this basket should be developed further.
Why creation of output tables in Oracle Data Miner is necessary?	Creation of the output tables speeds up the process of manipulations with data. Data Miner records the whole chain of procedures and starts every time the work with data from the beginning. The more SQL Queries are created, the longer is the process of data transformation. One step might take hours to be finalized. Output tables "break" this chain.
Is it a good idea to create "Women's basket"? It will contain cosmetics, clothes for women? Is this assumption worth checking?	At least this idea should be tested. Some filters can be created and used. Why not? Even some crazy ideas can lead to great results. This one is not even crazy, it is a very good assumption.
Why is it important to visualize market baskets? Oracle Data Miner identified a lot of cross-selling pairs.	Visualization can show the way more. Some hidden interconnections might not be visible. Visual models also simplify perception of the results. Apart from that, it is easy to

	communicate the results to other employees with different backgrounds. The output of the analysis will be used in other departments as well, so visualization will help in this case.
Market basket structures are obtained but there are also cross-selling pairs outside these baskets. They are represented on visual models and they are frequently purchased together. Will it be considered?	Yes of course. And it will be used. Separate report with cross-selling pairs will be created and sent to manager of the department. It is important not to underestimate the importance of other findings.
Why is it necessary to filter out top 10 products for analysis, for example, of "Alcohol basket"?	It is done to avoid additional intersections with other baskets. Intersections are still possible, however it is important to control the level of intersections. Market basket structures should be identified and they have to be different. This is the point.